Author:
Ward, Mary

Title:
Pathways to obesity

investigating the role of modifiable lifestyle factors through intermediate phenotypes

Pathways to obesity: investigating the role of modifiable lifestyle factors through intermediate phenotypes

Mary Elizabeth Ward

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences

Bristol Medical School

April 2018

Word count: 37,756

Abstract

The causal pathway between modifiable lifestyle factors and obesity is complex. The growing obesity epidemic impacting across the lifecourse is a major public health concern in many countries, therefore understanding the causes and consequences of adiposity is important. In this thesis I investigate the role of the metabolome and methylome in the relationship between dietary behaviour and obesity.

Establishing causality in observational studies is challenging due to unmeasured confounders and potential for reverse causation. I use Mendelian randomization (MR), two-sample MR and longitudinal analysis to infer causality in the relationships between diet, the methylome, the metabolome and body mass index (BMI). A good understanding of causality in these relationships is important to address the question of how the major public health problem of obesity should be tackled.

Dietary behaviour is a complex trait, and hence few studies have identified genetic variants associated with diet. I performed a GWAS of macronutrient intake in UK Biobank, with the aim of identifying genetic variants that could be used to generate a robust genetic instrument for dietary behaviour for use in MR.

Whilst studies have demonstrated the effect of adiposity on metabolic signatures from early adulthood onwards, there is a lack of published data exploring the relationship between adiposity and the metabolome in childhood. Using the Avon Longitudinal Study of Parents and Children (ALSPAC), I observed strong evidence of associations between BMI and several metabolite measures in childhood and adolescence in children, showing that the ability of BMI to influence the metabolome starts in childhood. Many BMI-associated metabolites are also associated with dietary behaviour, so it is likely that the metabolome plays a key role when trying to understand the relationship between dietary behaviour and BMI.

Several associations have been observed between BMI and methylation, mostly in adults. I investigated the relationship between BMI and methylation in childhood and adolescence and explored their relationship with dietary behaviour. My observations corroborated the prevailing evidence that DNA methylation occurs as a consequence (rather than a cause) of BMI.

Acknowledgements

There are several people who I would like to thank for their help in completing this project:

- My supervisors, Dr Tom Gaunt and Prof Caroline Relton, for their guidance and encouragement over the last few years.
- Wellcome, for funding this project.
- The ALSPAC participants and families, the midwives for their help in recruiting them, and the whole ALSPAC team, including interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.
- The UK Biobank participants.
- My colleagues at the MRC IEU and the Bristol Medical School, including Dr Rebecca Richmond and Dr Gemma Sharp with whom I did the *HIF3A* methylation and BMI work, Prof Nic Timpson for his feedback in my annual reviews, Dr Kate Northstone for her help with the diet PCs and Sharen O'Keefe for her administrative assistance.
- The many colleagues with whom I've shared an office over the last few years, for their moral support, advice and coffee breaks.
- My friends and family, for their endless encouragement, patience and support.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Table of contents

List of tables

List of figures

List of appendices

List of acronyms

2SLS	Two-stage least squares
ALSPAC	Avon Longitudinal Study of Parents and Children
ARIES	Accessible Resource for Integrated Epigenomics Studies
BMI	Body mass index
CLA	Conjugated linoleic acid
DXA	Dual-energy X-ray absorptiometry
EWAS	Epigenome-wide association study
FMI	Fat mass index
FFQ	Food frequency questionnaire
GCTA	Genome-wide complex trait analysis
GIANT	Genetic Investigation of Anthropometric Traits
GWAS	Genome-wide association study
HRC	Haplotype Reference Consortium
IV	Instrumental variable
IVW	Inverse-variance weighted
LD	Linkage disequilibrium
LIML	Limited information maximum likelihood
LOO	Leave-one-out
MAF	Minor allele frequency
mQTL	Methylation quantitative trait locus
MR	Mendelian randomization
MS	Mass spectrometry
MUFA	Monounsaturated fatty acid
NMR	Nuclear magnetic resonance
PC	Principal component
PCA	Principal components analysis
PUFA	Polyunsaturated fatty acid
RRR	Reduced rank regression
SNP	Single nucleotide polymorphism
WC	Waist circumference

CHAPTER 1. INTRODUCTION

1.1. Overview of the problem

The obesity epidemic is a major public health problem today. Obesity is associated with a range of comorbidities including cardiovascular disease, type 2 diabetes and some cancers.[1,2] People who are overweight or obese have lower life expectancies – studies have estimated reductions in life expectancy of 0-3 years, 1-6 years and 1-10 years for overweight, obese and very obese adults respectively, depending on the age and sex of the individual.[3] The number of healthy life-years lost is even greater. A 2011 study predicted that a rise in obesity-related diseases will cost the NHS nearly an extra £2 billion per year by 2030.[4]

Obesity prevalence is increasing in adults and children, both in the UK and globally.[2,5,6] The Health Survey for England 2015 (http://digital.nhs.uk/catalogue/PUB22610) found that 27% of adults in England were obese, and a further 31% of women and 41% of men were overweight. Two of the main causes of this increasing prevalence are thought to be increases in sedentary behaviour and the consumption of high-energy foods.[7]

Obesity as a public health issue starts in childhood, since children who are overweight have a greater risk of becoming overweight adults.[8] Although chronic diseases such as cardiovascular disease and type 2 diabetes are unlikely to have already developed in childhood, it is possible to study risk indicators such as hypertension and cholesterol levels or other metabolic perturbations.[9-11] Some adverse cardiovascular and metabolic features are already evident in children and young adults.[12,13]

A range of interventions have attempted to tackle childhood obesity including school-based interventions and family-based interventions; however, any success is often limited to the duration of the intervention.[14] Given that there has been a relative lack of success in developing and implementing interventions to address diet or physical activity to prevent obesity, there is therefore a strong motivation to increase our understanding of the intermediate pathways linking known risk factors with obesity. This may provide new intervention targets to prevent or treat obesity or ameliorate the consequences through altering intermediates along the pathway.

Obesity is a complex issue linked to a range of lifestyle traits and socioeconomic factors. Untangling the causal and molecular pathways to obesity is challenging and studies often lack power and are beset with confounding.

Population-based approaches are useful since obesity and overweight prevalence is high amongst the general population, and a wide range of small variations in lifestyle behaviours may be linked to excess adiposity.

Technologies have developed over the last few years, allowing for relatively low-cost measurement of high-dimensional molecular phenotypes ("omics" data) such as metabolites and DNA methylation. This omics data is now available, along with lifestyle and anthropometric data, in large cohort studies such as UK Biobank.[15] New statistical methods have also been developed which aim to attempt to deal with confounding, interrogate causality and investigate molecular mediation.[16,17]

This thesis aims to interrogate the relationship between dietary behaviour, molecular intermediates and adiposity. A clearer understanding of this relationship is key to informing future interventions, whether that be developing novel therapeutic or lifestyle interventions or advising public health policy.

1.2. Diet and adiposity

Both cross-sectional and longitudinal studies have investigated the relationship between dietary habits and BMI or other measures of obesity in children and adolescents. Findings from these studies include evidence from a systematic review suggesting that dietary energy density is positively associated with increased adiposity in children and adolescents.[18] A separate childhood study of macronutrient intake and body fat percentage provided more detail by showing that adiposity was positively associated with percentage of energy derived from fat and negatively associated with percentage of energy derived from carbohydrate.[19]

A study of dietary patterns and change in fat and lean mass observed that a diet high in fruit and vegetables but low in processed food was associated with a decrease in fat mass gain in girls, and, perhaps surprisingly, a diet high in sandwiches and snacks was also associated with a decrease in fat mass gain in girls and an increase in lean mass gain in boys.[20] A study of fast food consumption found that teenagers who consume fast food more frequently tend to eat less fruit and vegetables and have higher BMIs.[21]

Several studies of diet and adiposity have also been conducted in adults. A cross-sectional study of healthy older men has linked obesity to energy intake from fat.[22] Other studies have observed associations between higher consumption of meat, refined grains, sweets and desserts and long-term weight gain.[23,24] Changing dietary behaviour may have a positive effect on obesity – longitudinal studies of adults have found that positive changes in eating behaviour were accompanied by a decrease in BMI or a smaller weight gain.[25,26]

1.3. Dietary behaviour

Dietary behaviour is a complex trait. Assessing diet can be costly, and assessment methods suffer from a high degree of measurement error. There are many different aspects to dietary behaviour, and hence summarising dietary behaviour in a form that can be used in quantitative analysis can also be challenging.

1.3.1. Assessing dietary behaviour in cohort studies

Cohort studies typically assess dietary behaviour using food frequency questionnaires (FFQs),[19,22,23,25,27-30] food diaries,[31-33] or 24-hour recalls.[34,35] Studies have also used a range of other questions such as how frequently the participant skips breakfast or visits fast food outlets.[21,36]

FFQs are used to estimate a person's usual intake of a set list of foods, often more than 100 items long, which aims to encompass most of their diet.[37] FFQs commonly ask about usual food intake over the past year, though some FFQs may stipulate a shorter time period such as a month. Questions are usually multiple choice (e.g. eat a food item more

than once a day, 4-7 times a week, 1-3 times a week, once a fortnight, or never/rarely),[27] but some questions may require a number for an answer, for example average number of units of alcohol per week.

A diet diary, or food record, is a record of all food and drink consumed during a set period, typically between one and seven days in length.[37] Ideally, participants should record their food and drink intake at the time at which they consume it, or at least on the same day, to avoid relying too much on memory. Participants are asked to include information on portion sizes. More detailed diet diaries may also include information on food brands.

In a 24-hour diet recall, participants are asked to record their food and drink consumption from the last 24 hours. Data is typically collected during a face-to-face or phone interview with a field worker or nutritionist or via a structured web-based questionnaire.

Since FFQs capture usual food intake, they are less affected by one-off events such as holidays or illness. Diet diaries may be less representative of usual food intake since they depend on the particular days over which the diary was completed, and hence are vulnerable to any deviation from normal food intake.

FFQs tend to suffer from a considerable amount of measurement error due to a lack of detailed information on portion sizes.[37,38] People may also find it hard to accurately report how often they eat different foods. Studies comparing FFQs, diet diaries and 24-hour recalls have found that FFQs tend to record higher daily intakes of vegetables.[39] Diet diaries also suffer from measurement error, usually resulting in underestimation of total energy intake.[40] People with higher BMIs are more likely to underreport their energy intake.[41]

Dietary assessment methods are vulnerable to social desirability bias, for example a downward bias in overall reported food intake and an upward bias in reported fruit and vegetable intake.[42,43] Food diaries and scheduled 24-hour recalls are also susceptible to

reactivity bias, where a participant alters their eating behaviour either to make it simpler to record the foods and quantities consumed or for social desirability reasons.[37]

1.3.2. Summary variables for dietary behaviour

FFQs, diet diaries and 24-hour diet recalls collect data on great numbers of different food items. It is often helpful to summarize the data to reduce the number of variables to analyse ("dimensionality reduction"). One approach is to use empirical methods, such as principal components analysis (PCA) or cluster analysis, to identify dietary patterns.[44] Such methods have been used in this thesis and are discussed in greater detail in Chapter 2 (Methods) and in their application later in the thesis. Alternatively, studies may wish to derive new composite variables that estimate different aspects of dietary intake, such as macronutrient intake or fast food intake.[35,45]

1.3.3. Heritability of diet

Family and twin studies have estimated that genetic effects typically account for about 20% to 40% of variation in energy and macronutrient intake.[46] However, as of yet, few studies have identified genetic variants associated with dietary patterns or behaviours, often due to the paucity of high quality specific data at scale that accurately reflects dietary patterns and behaviours. Heritability is a very useful parameter as it provides useful information regarding the determinants of variation in a phenotype. Identification of genetic loci associated with a trait however, requires a "clean", unambiguous phenotype to be defined, which has posed a challenge in the context of diet. Functional insights can also be gained from characterising genetic loci that contribute to heritability. In addition, the use of genetic variation is hugely valuable in the application of causal analysis methods to strengthen inferences that can be made regarding observational associations which are often biased. Thus, the application of genetic variants in causal analysis methods is a strong motivating factor in identifying diet-related genetic variation.

1.4. BMI and other measures of adiposity

Body mass index (BMI), defined as weight (kg) divided by height squared (m^2), is the most commonly used measure of adiposity. The categories usually used in adults are: underweight, <18.5 kg/m^2; normal weight 18.5-24.9 kg/m^2; overweight 25-29.9 kg/m^2; and obese ≥30 kg/m^2.[2]

The BMI categories used in adults (for underweight, normal weight, overweight and obese) are not appropriate for use in children and adolescents since they have differing body proportions. For example, a 5-year-old and a 15-year-old may have the same BMI, but one could be considered a normal weight and the other overweight. To address this issue, age is often taken account of when studying BMI in children and adolescents, and reference charts are used to compare children against reference percentiles for their age.[47,48] This is called "BMI for age".

Studies of BMI in adults and BMI for age in children have found that these measures have a high specificity (low false positive rate) for identifying excess adiposity, but a low to moderate sensitivity (moderate to high false negative rate).[48,49]

Other easy-to-measure anthropometric measures of adiposity include variations on waist circumference, e.g. waist-hip ratio and waist-height ratio. Compared to BMI, waist circumference captures central adiposity better, but is a little less frequently measured and does not reflect general adiposity as well as BMI.

A study comparing childhood BMI with childhood waist circumference and childhood fat mass as predictors of cardiovascular risk factors in adolescence found that all three childhood adiposity measures were similarly capable of predicting an adverse cardiovascular profile in adolescence.[50] BMI is the primary focus of this thesis because of the ubiquity of BMI as a variable.

1.4.1. Heritability of BMI

A review of the heritability of BMI found that BMI heritability estimates from twin studies ranged from 0.47 to 0.90 (median=0.75).[51] Heritability estimates were 0.07 higher (p=0.001) in children than adults; and increased with mean age in childhood studies but decreased with mean age in adult studies. A genome-wide association study (GWAS) of ~320,000 adults identified 97 BMI-associated genetic loci, accounting for ~2.7% of variation in BMI.[52] More recently, a larger BMI GWAS has been published (n~700,000) which identified 941 BMI-associated genetic loci, accounting for ~6.0% of variation in BMI.[53] This genetic variation can be leveraged in Mendelian randomization (MR) analysis to understand the consequences of variation in BMI.[54,55]

1.5. Mediating mechanisms

Whilst dietary behaviour and adiposity are known to be related, less is known about the role of molecular intermediates in this relationship. Potential intermediates include various “omics”, for example the epigenome, metabolome and gene products such as transcriptome and proteome.[17]

This thesis aims to investigate the role of the metabolome and methylome in the relationship between dietary behaviour and adiposity. Data for these intermediates has recently become available in relatively large population samples, and hence analyses in this thesis were able to exploit existing data. Proteomics data are now becoming more widely available, but were less so when this project commenced, and the availability of transcriptomics data is still quite unusual in large population cohorts because of the difficulties of sampling.

1.5.1. The metabolome

The metabolome is the collection of all small molecules (metabolites) in a cell or tissue that are involved in metabolic reactions and are needed for the growth, maintenance and normal function of the cell.[56] Metabolite profiles have been studied in relation to a

range of traits and diseases, for example physical activity, hypertension and type 2 diabetes.[57]

The two main technologies used to measure the metabolite profile are nuclear magnetic resonance (NMR) and mass spectrometry (MS).[57] Both technologies provide extensive metabolite coverage, though the coverage provided by MS is more extensive. However, NMR is cheaper and is therefore generally more suitable for large cohort studies. Additionally, NMR can analyse lipoproteins, but MS cannot.

1.5.1.1 Heritability of the metabolome

Kettunen et al. studied the heritability of metabolite measures assayed by NMR in young adults from the Finnish Twin Cohort.[58] Their heritability estimates ranged between 0.48-0.62 for lipids, 0.50-0.76 for lipoproteins, and 0.23-0.55 for amino acids and other small-molecule metabolites. As outlined above for adiposity, insights in to the genetic determination of metabolite traits can be useful in understanding molecular pathways and in fuelling MR analyses to strengthen causal inference.

1.5.1.2 Diet and the metabolome

Many studies have observed associations between dietary patterns and blood metabolite profiles.[59-65] These findings include an association between a Western diet and higher levels of amino acids,[59] and positive associations between fruit and vegetable intake and phosphatidylcholines.[63] Some studies have looked at metabolites individually or by class,[60,65] whilst other studies used PCA to summarise metabolite profiles.[59,61]

It is difficult to accurately assess dietary intake in populations, therefore it would be hugely beneficial if a clearer understanding of relationship between diet and metabolites could be used to accurately index an individual's dietary pattern given their metabolic profile.[64,66,67]

1.5.1.3 BMI and the metabolome

Several studies have observed strong associations between adiposity and the human serum metabolome.[68-78] These include studies in children which have observed associations between obesity status and amino acid levels,[68,69] and studies of young adults which have observed strong links between adiposity and lipoproteins, amino acids and fatty acids.[70,71] A young adult study of several obesity measures (including waist circumference, android fat (%) and subcutaneous fat) found that abdominal fat was overall most strongly associated with an adverse metabolite profile.[71] Findings from studies of (or including) middle-aged and older adults include a study in women which found that obese women had significantly higher branched-chain amino acid levels than lean or overweight women.[72]

Some studies have also tried to infer causality in the relationship between BMI and the metabolome.[70,75,76] A study of young adults which found that elevated BMI was associated with adverse changes in the metabolite profile also conducted MR analysis.[54,70] Their results suggest that adiposity has a causal effect on the metabolic profile. They also observed that change in BMI was associated with changes in the metabolite profile, suggesting that the metabolite profile can be modified through lifestyle changes. Two other studies in adults investigating the causal effect of BMI on metabolic traits observed evidence suggesting that BMI has a causal effect on HDL cholesterol levels.[75,76] The literature to date in this area is largely (but not exclusively) confined to adults, with little evidence of the interrelationship of emerging adiposity with the metabolome in children.

1.5.2. The methylome

DNA methylation in humans is an epigenetic modification of DNA in which methyl groups attach to CpG dinucleotides.[79,80] Methylation patterns vary between individuals, are tissue-specific, mitotically stable, and can change in response to lifestyle factors, for example smoking behaviour.[81-83] Methylation profiles in large cohort studies are usually assayed using arrays such as the Illumina 450K HumanMethylation BeadChip array.[84] An (epi)genome-wide approach has been widely adopted, following the example of the

GWAS approach. Testing associations of large numbers of methylation sites for evidence of association with a specific exposure or phenotype is now routine and has been termed an epigenome-wide association study (EWAS).

1.5.2.1 Diet and the methylome

Identifying associations between dietary behaviour and DNA methylation has proved challenging, and few epigenome-wide association studies (EWAS) of diet have been able to identify robust associations.[85,86] This is likely due to small sample sizes and poor specificity of dietary measures.

1.5.2.2 BMI and the methylome

The relationship between adiposity and DNA methylation has been investigated in various large-scale studies.[87-93] The first major EWAS to report robust associations between CpG sites and BMI was a study in adults by Dick et al. which identified associations between increased methylation at three CpGs in *HIF3A* and increased BMI.[91] This finding was further explored as part of this thesis and is described in more detail in Chapter 7.

Two of the largest EWAS to date were performed by Wahl et al. and Mendelson et al.[94,95] Wahl et al. performed an EWAS of BMI in 10,261 adults and identified 187 CpGs associated with BMI at an epigenome-wide level (defined as $p < 1 \times 10^{-7}$ here).[94] The genetic loci identified by these 187 CpGs include genes involved in lipid and lipoprotein metabolism. Mendelson et al. performed a BMI EWAS in 7,798 adults and identified 83 BMI-associated CpGs.[95] 38 BMI-associated CpGs were common to both EWAS.

Wahl et al. and Mendelson et al. both conducted MR to investigate causality between BMI and their BMI-associated CpGs. Both studies agreed that changes in methylation mostly appear to be a consequence of changes in BMI, rather than a cause.

1.6. Novel approaches to understanding pathways between diet and BMI

Pre-omics approaches to understanding the causality between dietary behaviour and adiposity have included intervention studies and longitudinal studies. The omics era has led to the development of novel approaches to understanding pathways between dietary behaviour and adiposity.

1.6.1. Cohort resources

The scope to interrogate the relationship between diet and adiposity has developed enormously over recent years due to the increasing availability of large scale population level data. Not only genotype data, but reliable phenotype data with dietary factors measured at scale, create new opportunities to explore the pathways between diet and BMI. Effect sizes are often small when studying complex lifestyle-related phenotypes, and large sample sizes are needed to identify these small effects. Large cohorts with omics data such as UK Biobank and ALSPAC are key to understanding these relationships.[15,96]

1.6.2. Molecular phenotyping

The development and adoption of robust platforms for high throughput analysis of molecular phenotypes including metabolites, DNA methylation profiles etc, have facilitated the application of epidemiological methods to gain insights in to the pathways of interest here. Genome-wide association studies (GWAS) are used to identify genetic variants associated with phenotypes and gain better understanding of genetic influences. These genetic variants can be used to create genetic instruments for use in MR analysis to explore causality between phenotypes.[54,55] In a similar way, EWAS can be used to identify differentially methylated CpGs across the epigenome that are associated with diet and BMI; and metabolome-wide association studies can identify metabolites associated with both diet and BMI.

1.6.3. Causal inference methods

The methodological development of Mendelian randomization (MR) as a tool to strengthen causal inference has escalated rapidly over recent years with widespread adoption of the method. Furthermore, methodological refinement of statistical methods for molecular mediation provides new opportunities to produce robust evidence. MR is a form of instrumental variable analysis that uses genetic variants as instruments for the exposure of interest.[54,55] These genetic variants should be independent of the outcome given the exposure, and independent of any confounders of the exposure-outcome relationship.

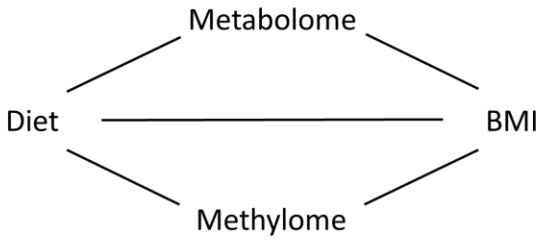
1.7. Summary

In summary, obesity is a major public health problem with several comorbidities, and prevalence of overweight and obesity is increasing. Dietary behaviour is strongly linked to excess adiposity; however, this relationship is not fully understood and dietary interventions often have limited success. Large cohort studies have collected genetic, epigenetic and metabolite data along with measures of dietary behaviour and adiposity. Statistical methods including causal inference methods and mediation analysis may be used to investigate the relationship between diet, adiposity and molecular intermediates. The recent growth in data availability has only now made it possible to systematically and comprehensively analyse the role of these molecular intermediates. In summary, through the application of state-of the-art epidemiological methods the aim of this thesis is to gain greater insights into the pathways linking diet and BMI with a view to enhancing the evidence base for future prevention and treatment of obesity and its comorbidities.

1.8. Overarching aims of thesis

The main aims of this thesis are to use statistical methods to investigate the role of the metabolome and methylome in the relationship between dietary behaviour and obesity. The primary hypotheses proposed are that metabolites and/or DNA methylation are intermediate traits in the relationship between diet and BMI (**Figure 1**). A key method used to assess causality will be MR, for which genetic instruments are needed. Although genetic instruments have been found for BMI, there are few known genetic variants robustly associated with dietary behaviour, and hence in Chapter 3 I perform a GWAS of dietary intake with the aim of identifying suitable genetic instruments for diet. In Chapter 4 I investigate the relationship between dietary behaviour and BMI in UK Biobank and ALSPAC. In Chapter 5 I investigate the relationship between BMI and the metabolome. In Chapter 6 I explore the association between diet and the metabolome, and the role that the metabolome plays in the relationship between diet and BMI. In Chapter 7 I explore the relationship between BMI, methylation and diet.

Figure 1 – Framework representing the hypothesis explored in this thesis.



CHAPTER 2. METHODS

Analyses in this thesis use data from two UK cohorts: ALSPAC, a longitudinal birth cohort;[96,97] and UK Biobank, a population-based prospective cohort.[15,98,99] This chapter describes these cohorts and the variables from them that are used in this thesis. This chapter also describes the main statistical methods applied in this thesis: linear regression, GWAS, EWAS, Mendelian randomization and mediation.

2.1. Data sources

This section describes how the data used in this thesis were generated by other researchers.

2.1.1. Avon Longitudinal Study of Parents and Children

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a longitudinal birth cohort study that recruited expectant mothers of 14,541 pregnancies with due dates between 1st April 1991 and 31st December 1992 living in the former county of Avon, UK.[97] Avon was made up of what is now Bristol and parts of North Somerset and South Gloucestershire. The catchment area is comprised of three NHS District Health Authorities (DHAs) – Southmead DHA; Frenchay DHA; and Bristol and Weston DHA. 14,062 children were live-born from these pregnancies, of whom 13,988 children were alive at 1 year of age.[96] When the children were 7 years old a further recruitment drive was done, resulting in an additional 452 children being enrolled. Between the ages of 8 and 18 years another 257 children were enrolled.

Data has been collected on the children and their mothers in various ways including questionnaires and clinical assessments. Data was collected on the children at 68 data collection timepoints between birth and age 18 years: 25 questionnaires about the children completed by the mothers or main caregivers; 34 questionnaires that the children completed about themselves; and 9 "Focus" clinics.[96] Phenotypes covered by the questionnaires include health, developmental, psychological and social measures. The 9 Focus clinics held between the age of 7 and 17 years collected physiological, cognitive, psychological and social measures. Blood samples were taken at the clinics,

from which genetic, epigenetic and metabolomic measures have been generated. Data has also been collected from education questionnaires and assessments administered in the children's schools.

Data has been collected on the mothers from questionnaires (18 questionnaires administered between pregnancy and 20 years postnatal), medical records (obstetric data) and 4 clinical assessments (held between 2008 and 2015).[97]

Data from the 1991 census has been used to compare socio-demographic characteristics of mothers in Great Britain with those of ALSPAC mothers from the 8-month postnatal questionnaire.[97] These comparisons found that ALSPAC mothers are more likely to live in owner-occupied accommodation, have a car in the household, be married, and less likely to be non-White.

As is common in longitudinal studies, ALSPAC has experienced some attrition over the years.[96] At least one data item was completed for 11,408 children during the "late childhood" phase (>7 and <13 years of age); 9,600 children during the "adolescence" phase (≥13 and ≤16 years); and 7,729 children during the "transition to adulthood" phase (>16 and ≤18 years).

A timeline showing the data collection timepoints of the children's diet, adiposity, metabolite, and methylation measures studied in this thesis is found in **Figure 2**.

Figure 2 – Timeline of the data collection timepoints of the children's diet, adiposity, metabolite, and methylation measures studied in this thesis.

Numbers refer to approximate age in years. *Methylation data is available from blood samples taken at either 15 years or 17 years; data from these two ages is combined to create a single timepoint, mean age 17.1 years.

2.1.1.1 Diet

Food frequency questionnaires

Dietary intake has been measured in the ALSPAC children using food frequency questionnaires (FFQs) at ages 2, 3, 4, 7, 9 and 13 years and diet diaries at ages 5, 7, 10 and 13 years.[100] For the analyses in this thesis, only data collected at 7 and 13 years were used since these were the closest ages to when methylation and metabolite measures were also available (7 years, and 15 and/or 17 years).

When their child was ~7 years old the mothers (or main carers) of the ALSPAC children were asked to complete a FFQ, covering 57 different food types.[27] The FFQ asked the parents how often their child consumed each food "nowadays". The questionnaire covered food provided by the parents, but not food provided by others outside the home such as school dinners. 8,515 questionnaires were returned, with a mean age at completion of 6.85 years.

When the children/teenagers were ~13 years old, two distinct FFQs were sent to the mothers and teenagers.[101] The mother's FFQ asked her to record how often her teenager consumed different foods "nowadays", but to only include food that she provided for her teenager (including packed lunches but not school dinners). The teenager's FFQ asked them to record how often they consumed various different foods "nowadays" that were not covered by the mother's FFQ, for example school dinners or foods bought outside of school and additional snacks and drinks. The foods from the mother's and teenager's FFQs were combined to make 62 food groups. The teenager's and mother's FFQs were both returned for 6,203 teenagers, at mean age 13.1 years.

Principal components analysis (PCA) has been performed to identify dietary patterns in the FFQs at age 7 years and at age 13 years. For the FFQ at age 7 three principal components (PCs) emerged which best described the children's dietary patterns, these are "junk", "traditional" and "health conscious".[27] The "junk" PC is mainly associated with foods with a high fat and sugar content and processed foods; the "health conscious" PC with vegetarian foods, rice, pasta, salad and fruit; and the "traditional" PC

with a traditional British diet based on meat, potatoes and vegetables. This PC data is available for 8,286 children.

Four dietary PCs were derived from the dual-source FFQ at age 13 years.[101] The "traditional/health conscious" PC is associated with higher intakes of meat, fish, eggs, rice, pasta, salad, vegetables and pulses. The "processed" PC is associated with higher intakes of processed food such as processed meat, coated chicken and fish products, pizza and chips. The "snacks/sugared drinks" PC is associated with higher intakes of crisps, biscuits, chocolate, sweets, squash and fizzy drinks. The "vegetarian" PC is associated with higher intakes of meat substitutes, nuts and pulses. These PCs are available for 5,418 children.

Diet diaries

The children were invited to attend a research clinic when they were 7 years old. The mothers were sent a 3-day diet diary to complete prior to the visit, recording all food and drink consumed by their child over two weekdays and one weekend day.[31,102]

The children/teenagers were also invited to attend a clinic when they were 13 years old. Prior to this visit the teenagers were sent a 3-day diet diary to complete themselves, recording their food and drink consumption across two weekdays and one weekend day.[31,100] At the clinic a trained nutrition fieldworker interviewed the teenager and accompanying parent to clarify any uncertainties.

PCA has been performed to identify dietary patterns in the data from the diet diaries at ages 7 years and 13 years (Northstone et al., unpublished), using the same method that is described for the 10-year-olds' by Smith et al.[103] The three PCs derived from the 7-year-olds' diet diaries were named the "health aware", "traditional" and "packed lunch" PCs. The "health aware" PC is associated with higher intakes of cheese, high fibre bread, pasta, salad, fresh fruit and fruit juice, and lower intakes of processed meat, chips and diet fizzy drinks. The "traditional" PC is associated with higher intakes of poultry, red meat, vegetables and roast potatoes, and a lower intake of chips. The "packed lunch" PC

is associated with higher intakes of low fibre bread, margarine, ham, bacon, crisps and diet squash.

The three PCs derived from the 13-year-olds' diet diaries were also named "health aware", "traditional" and "packed lunch". The "health aware" PC is associated with higher intakes of cheese, yoghurt, high fibre bread, breakfast cereal, pasta, salad, legumes, nuts, fresh fruit and water, and lower intakes of coated and fried chicken, processed meat, chips, fizzy drinks and diet fizzy drinks. The "traditional" PC is associated with higher intakes of vegetables and roast potatoes, and lower intakes of chips and salad. The "packed lunch" PC is associated with higher intakes of low fibre bread, margarine, ham, bacon, sugar, biscuits, crisps, water, diet squash, tea and coffee, and a lower intake of rice.

Motivation for using these diet summary variables

Both PCA and cluster analysis have been used to summarise the main patterns in the FFQ data.[44,101] PCA was used to study correlations between the different food groups measured by the FFQ and identify linear combinations (PCs) of foods that are often consumed together. Cluster analysis was used to group the participants into non-overlapping clusters according to similarities in their diets. Both methods identified three main dietary patterns, and there are strong similarities between the patterns identified by PCA and those identified by cluster analysis.[44] For the analyses in this thesis, PCs are used since they are continuous measures, unlike the clusters which are discrete.

PCA has also been used to summarise the main patterns in the diet diary data.[31,103] Food item data from diet diaries needs to be quantified to perform PCA, and the data is commonly quantified by food weight or as binary variables. A study of the effect of different forms of input variable quantification on diet diary data from the ALSPAC 10-year-olds concluded that PCs generated using food weight data were more interpretable.[103] Hence, for the analyses in this thesis, diet diary data was quantified according to the estimated weight (in grams) of each food consumed, before being summarised using PCA.

2.1.1.2 BMI/adiposity

The children's height and weight were measured at the research clinics at ages 7, 9, 13, 15 and 17 years. Height was measured to the last complete mm using a Harpenden Stadiometer. Weight was measured to the nearest 0.1kg using Tanita scales. These measures were used to calculate the child's BMI (kg/m^2) at each age.

At the age 9, 13, 15 and 17 years clinics, fat mass (kg) was measured using a Lunar Prodigy dual-energy X-ray absorptiometry (DXA) scanner. Fat mass index (FMI) (kg/m^2) was calculated as fat mass (kg) divided by height squared (m^2) at each age.

2.1.1.3 Covariates and other variables of interest

The mothers were asked about their highest educational qualification in a questionnaire administered during pregnancy. This information has been categorized as a binary variable of whether the mothers completed A-levels/a university degree or not. In this thesis, maternal education is used as a proxy for socioeconomic status.

Teenagers were asked about their smoking status in questionnaires administered at the age 15 years and age 17 years clinics. This information has been categorized as never/less than weekly, weekly, or daily.

2.1.1.4 Genotyping

Biological samples including blood samples for DNA isolation have been collected for approximately 10,000 ALSPAC children, from which genome-wide SNP data has been generated for >8,000 children. 9,912 ALSPAC children were genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform by 23andMe, who subcontracted the Wellcome Trust Sanger Institute (Cambridge, UK) and the Laboratory Corporation of America (Burlington, NC, US).[104] Individuals were excluded if they had incorrect sex assignments; minimal or excessive heterozygosity (<0.32 and >0.345 for the Sanger data and <0.31 and >0.33 for the LabCorp data); disproportionate levels of individual missingness (>3%); evidence of cryptic relatedness (>10% IBD). The remaining individuals were assessed for evidence population stratification by multidimensional

scaling (MDS) analysis and compared with HapMap (release 22) European descent, Han Chinese, Japanese and Yoruba reference populations; and all individuals of non-European ancestry were excluded to avoid population stratification.[105] In total 1,547 children were excluded leaving a sample size of 8,365 children. SNPs were removed if they had a minor allele frequency (MAF) <1%; a call rate of <95%; or if they were not in Hardy-Weinburg equilibrium. The genotype data was imputed to the 1000 Genomes reference panel (Version 1, Phase 3).[106]

2.1.1.5 Methylation

The Accessible Resource for Integrated Epigenomics Studies (ARIES) project is a sub-study of ALSPAC.[107] A subset of 1018 mother-offspring pairs were selected from ALSPAC based on the availability of DNA samples from three time-points for the children (birth, childhood c.7 years old and adolescence c.15-17 years old) and two time-points for the mothers (during pregnancy and c.15-17 years later). The DNA samples taken at birth were extracted from cord blood, whereas the childhood and adolescent DNA samples were extracted from peripheral blood. The childhood DNA sample was taken at the 7-year-old ALSPAC clinic visit (mean age 7.5 years) and the adolescent sample was taken at either the 15-year-old or 17-year-old clinic visit (mean age 17.1 years).

DNA methylation was quantified using the Illumina Infinium HumanMethylation450K BeadChip assay.[108] The assay measures the proportion of molecules methylated at each CpG site featured on the array. The methylation level at each CpG was calculated as a β-value, which is the ratio of the methylated probe intensity to the overall intensity and ranges from 0 (no cytosine methylation) to 1 (complete cytosine methylation).[109]

2.1.1.6 Metabolites

Metabolite profiles for ALSPAC participants have been generated from serum samples taken at the 7, 15 and 17 year clinics.[110] Metabolite profiles are available from at least one clinic for 7176 participants, of whom 1453 have metabolite profiles from all three clinics. Samples from the 7 year clinic are non-fasting samples, whereas samples from the 15 and 17 year clinics are fasting samples. Fasting samples taken in the morning

followed an overnight fast. Fasting samples taken after 2pm followed a fast of least 6 hours.

Metabolite measures (n ≤ 233) were quantified using a high-throughput proton (^{1}H) serum nuclear magnetic resonance (NMR) platform.[57,110] The metabolites measured by the platform and used in analyses in this thesis are listed in **Table 1**; these include lipoprotein lipids and subclasses, glycerides, phospholipids, fatty acids, amino acids and glycolysis-related metabolites.

NMR data were measured for three molecular windows: lipoprotein lipids (LIPO) and low molecular-weight metabolites (LMWM) which are acquired from native serum, and lipid extracts (LIPID) which is acquired from serum lipid extracts.[111]

Before use, the serum samples were stored at -80°C.[111,112] They were then thawed slowly overnight, before being prepared in a Gilson Liquid Handler 215 which performs automated sample preparation to 5mm outer-diameter NMR tubes, in which 300µl of sodium phosphate buffer are mixed with 300µl of serum. The prepared samples were put stored in 96-tube racks which were inserted into the robotic sample changer. The NMR data for the LIPO and LMWM windows were measured using a Bruker AVANCE III spectrometer operated at 500MHz. After these measurements, lipid extraction was performed and the NMR data for the LIPID window was measured using a Bruker AVANCE III spectrometer operated at 600MHz.

Table 1 – Metabolite measures.

VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein; HDL, high density lipoprotein.

Category	Name/subtype
Chylomicrons and extremely large VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Very large VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Large VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Medium VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Small VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Very small VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
IDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Large LDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Medium LDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Small LDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Very large HDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Large HDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Medium HDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Small HDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Lipoprotein particle sizes	VLDL particle size
	LDL particle size
	HDL particle size
Cholesterol	Total cholesterol
	VLDL cholesterol
	Remnant cholesterol
	LDL cholesterol
	HDL cholesterol
	HDL2 cholesterol
	HDL3 cholesterol
	Esterified cholesterol
	Free cholesterol

Glycerides & phospholipids	Triglycerides
	VLDL triglycerides
	LDL triglycerides
	HDL triglycerides
	Diacylglycerol
	Ratio of diacylglycerol to triglycerides
	Phosphoglycerides
	Ratio of triglycerides to phophoglycerides
	Phosphatidylcholine and other cholines
	Total cholines
Apolipoproteins	ApoA-I
	ApoB
	ApoB/ApoA-I
Fatty acids & saturation	Total fatty acids (FA)
	Estimated fatty acid chain length
	Estimated degree of unsaturation
	Docosahexaenoic acids (DHA)
	Linoleic acid (LA)
	Conjugated linoleic acid (CLA)
	Omega-3 fatty acids
	Omega-6 fatty acids
	Polyunsaturated fatty acids (PUFA)
	Monounsaturated fatty acids (MUFA)
	Saturated fatty acids (SFA)
	DHA to total FAs ratio
	LA to total FAs ratio
	CLA to total FAs ratio
	Omega-3 to total FAs ratio
	Omega-6 to total FAs ratio
	PUFAs to total FAs ratio
	MUFAs to total FAs ratio
	SFAs to total FAs ratio
Glycolysis related metabolites	Glucose
	Lactate
	Pyruvate
	Citrate
Amino acids	Alanine
	Glutamine
	Histidine
	Isoleucine
	Leucine
	Valine
	Phenylalanine
	Tyrosine
Ketone bodies	Acetate
	Acetoacetate
	3-hydroxybutyrate
Fluid balance	Creatinine
	Albumin (signal area)
Inflammation	Glycoprotein acetyls

2.1.2. UK Biobank cohort

UK Biobank is a population-based prospective cohort of ~500,000 participants who were recruited from across the UK between 2006 and 2010.[15,98,99] Participants were aged 40-69 years at recruitment. Participants were required to make a baseline visit to one of 22 assessment centres. At this initial assessment visit various data were collected by means of a questionnaire and a computer-assisted interview, including sociodemographic, lifestyle and health status data; several physical measures were also assessed, including anthropometric measures.[15] Blood samples were also collected at this assessment visit, allowing for genotype data to be assayed.[113]

2.1.2.1 Genotype data

UK Biobank participants were genotyped using the UK BiLEVE array or the UK Biobank axiom array, which contain ~800,000 markers.[113] Genotype data was imputed to the Haplotype Reference Consortium (HRC) reference panel (~40 million SNPs, of which ~11 million SNPs remain after filtering).[114]

Quality control filtering of the UK Biobank data was conducted by R.Mitchell, G.Hemani, T.Dudding, L.Paternoster as described in the published protocol (doi:10.5523/bris.3074krb6t2frj29yh2b03x3wxj).[115] Individuals whose reported sex did not match their genetic sex were excluded, as were those with sex-chromosome aneuploidy. The sample was restricted to individuals of white British ancestry who described themselves as "White British" and who have similar ancestral backgrounds.[113] Related individuals were identified using the KING toolset and removed.[116] SNPs were restricted to autosomal SNPs within the HRC site list.[114]

2.1.2.2 Diet data

Dietary data in UK Biobank was collected using the Oxford WebQ – a web-based 24-hour dietary recall questionnaire (http://biobank.ctsu.ox.ac.uk/crystal/docs/DietWebQ.pdf).[117] Since the questionnaire is web-based, it provides a low-cost method for assessing dietary intake in large cohort

studies. The questionnaire captures data on a person's food and drink intake the previous day, including portion sizes. This data is then used to calculate nutrient estimates automatically.

The questionnaire was first introduced towards the end of the recruitment period, hence diet data from the assessment visit is only available for the last ~70,000 participants recruited. All participants (who had provided UK Biobank with an email address) were invited, via email, to complete the questionnaire at four later occasions between February 2011 and June 2012. The dietary data was used to estimate intakes of various macronutrients during each 24-hour recall period. This thesis looks at 8 energy and macronutrient intake estimates in UK Biobank, which are listed in **Table 2**.

2.1.2.3 BMI measurement

Participants' height and weight were measured at the baseline assessment visit (http://biobank.ctsu.ox.ac.uk/crystal/docs/Anthropometry.pdf). Height was measured using a Seca 240cm height measure. Weight (kg) was measured using a Tanita body composition analyser. BMI (kg/m^2) was calculated as weight (kg) divided by height squared (m^2).

Table 2 – UK Biobank energy and macronutrients studied in this thesis.
Data-field numbers correspond to the UK Biobank Data Showcase (http://biobank.ctsu.ox.ac.uk/crystal/).

Dietary intake estimate	Data-field
total energy (kJ/day)	100002
protein (g/day)	100003
total fat (g/day)	100004
carbohydrate (g/day)	100005
saturated fat (g/day)	100006
polyunsaturated fat (g/day)	100007
total sugars (g/day)	100008
Englyst dietary fibre[118] (g/day)	100009

2.2. Methods

This section describes the core methods I used in my research. Additional detail is provided in the methods section of analysis chapters.

2.2.1. Linear regression

Linear regression analyses were performed in R (version 3.3.3) using the *lm* function from the *stats* package to fit Ordinary Least Squares (OLS) regression models. For example, if sugar intake is the exposure variable, BMI is the outcome, and the covariates are age and sex, then the code to fit the linear regression model is

```
lm(BMI ~ sugar intake + age + sex)
```

2.2.2. GWAS

Genome-wide association studies (GWASs) are a method for identifying associations between traits and genetic variation in a study population. Genome-wide SNP data is assayed, and then statistical analysis is used to test the relationship between each of those SNPs and a given trait. The statistical power of a GWAS to detect associations with a trait depends on the sample size and the effect sizes and frequencies of the trait-associated SNPs.[119]

2.2.2.1 UK Biobank GWAS pipeline

GWAS were performed using a pipeline developed within the MRC-IEU developed by B.Elsworth, R.Mitchell, C.Raistrick, L.Paternoster, G.Hemani, T.Gaunt (doi: 10.5523/bris.2fahpksont1zi26xosyamqo8rr).[120]

Quality control of the genetic data was performed using the methods described above in **2.1.2.1**. Phenotype and covariate files were submitted to the pipeline using the pipeline's submission spreadsheet. GWAS were conducted using linear regression implemented in PLINK v2.00.[121] Models are adjusted for genotype array, sex and the first 10 genetic PCs.

2.2.3. EWAS

Epigenome-wide association studies (EWAS) are used to identify epigenetic variation (commonly DNA methylation) associated with a chosen trait.[122]

In this thesis, EWAS were performed in R (version 3.4.1) using the *meffil* R package (https://github.com/perishky/meffil/).[123] *meffil* provides a computationally-efficient approach for performing functional normalisation to separate biological variation from technical variation.

EWAS regression models were fitted using Independent Surrogate Variable Analysis (ISVA), which models confounding factors as statistically independent surrogate variables.[124] Analyses were conducted using methylation ß-values (**2.1.1.5**). Models were adjusted for cell counts (B-cells, CD4+ T-cells, CD8+ T-cells, granulocytes, monocytes and NK), which were estimated using a reference-free method developed by Houseman et al.[125] CpG sites located on the sex chromosomes were excluded from analyses. CpG sites with a high detection p-value (>0.05 for >5% of samples) were also excluded.

The *meffil.ewas* function from the *meffil* R package was used to test the association between the trait of interest and methylation ß-values at each CpG site. For example, if "meth" is a matrix of the methylation data, "pheno" is a vector of the trait data and "covars" is a matrix of the model covariates, the code used to perform the analysis is

```
meffil.ewas(beta=meth, variable=pheno, covariates=covars,
      winsorize.pct=NA, most.variable=min(nrow(meth),20000),
      outlier.iqr.factor=3)
```

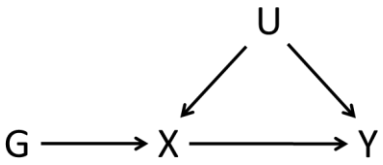
The R script used to conduct the EWAS using *meffil* was compiled by Dr Gemma Sharp.

2.2.4. Mendelian randomization

Mendelian randomization (MR) is a form of instrumental variable (IV) analysis that uses genetic variants as instruments.[54,55] MR is used to strengthen causal inference.

The following assumptions are made in MR (where G is the IV, X is the exposure, Y is the outcome, and U is the confounders):[55]

1. G is associated with X
2. G is independent of U
3. G is independent of Y given X



Several methods are available for conducting MR.[55] One of the most common methods used is the two-stage least squares (2SLS) method, which derives the causal estimate by first performing least-squares regression of the exposure variable on the IV(s), and then performing least-squares regression of the outcome variable on the predicted values from the first least-squares regression. Other common methods used to conduct MR include the limited information maximum likelihood (LIML) method.

Allele scores summarise multiple genetic variants associated with a trait as a single variable.[126] An allele score may be unweighted (the sum of the trait-increasing alleles) or weighted (using the genetic effect size estimates for each allele on the trait). The weighted BMI allele scores used in this thesis were created using the 97 SNPs and effect sizes from the Genetic Investigation of Anthropometric Traits (GIANT) consortium GWAS of BMI conducted by Locke et al. (n~320,000 adults).[52] Since the analyses in this thesis were performed, a larger BMI GWAS has been published (n~700,000 adults) which is a meta-analysis of the GIANT BMI GWAS and a GWAS in UK Biobank.[53]

A GWAS of BMI in ~48,000 children aged between 2 and 10 years identified 15 loci associated with childhood BMI.[127] In this thesis, the GIANT 97-SNP adult BMI allele score is the only allele score used to instrument BMI in analyses conducted in adults and also in analyses conducted in children and adolescents. An alternative to using the GIANT adult BMI score to instrument BMI in children and adolescents would be to use the

childhood-specific 15-SNP BMI score. Whilst the childhood-specific BMI score explains more variance in childhood BMI than the adult BMI score, it is not known which BMI score is more suitable for instrumenting BMI in adolescents. The decision was taken to use the adult BMI score throughout this thesis for more consistency between analyses conducted in children, adolescents and adults.

Burgess et al. conducted a simulation study to compare the use of an allele score (as the single genetic instrument in MR) with the use of multiple genetic instruments in conventional 2SLS and LIML methods.[126] They concluded that allele scores are suitable genetic instruments for MR if each of the genetic variants that make up the allele score satisfy the IV assumptions. They also concluded that allele scores allow greater numbers of genetic variants to be reliably used in an MR analysis than conventional 2SLS and LIML methods do. 2SLS and LIML give identical results when an allele score is used as the single genetic instrument in MR.

MR analyses were conducted in R (version 3.3.3) using the *ivreg* function from the *AER* package, which fit regression models using 2SLS. For example, to estimate the causal effect of BMI on diet, the following code could be used

```
ivreg(diet ~ BMI + age + sex | BMI allele score + age + sex)
```

2.2.4.1 Two-sample Mendelian randomization

Obtaining a large sample in which both the exposure and outcome traits are available can be challenging when studying traits such as dietary behaviour. Two-sample MR overcomes this issue by allowing the exposure and outcome to be measured in separate samples.[16,128] Two-sample MR takes instrument-exposure coefficients from one sample and instrument-outcome coefficients from a separate sample and uses these coefficients to calculate the MR estimates. These coefficients are often taken from publicly-available GWAS summary data.

In cases where the outcome trait is a continuous variable, methods commonly used to perform two-sample MR include inverse-variance weighted (IVW) regression and the

Wald ratio method. The IVW method calculates the MR estimate by combining the ratio estimates of the causal effects from each genetic variant using an inverse-variance weighted fixed-effect meta-analysis.[129,130] The Wald ratio method is used when there is only a single genetic instrument available.[55,128] The Wald ratio MR estimate is calculated by dividing the instrument-outcome coefficient by the instrument-exposure coefficient.

Two-sample MR analyses undertaken in this thesis were conducted in R (version 3.3.3) using the *mr_singlesnp* function from the *TwoSampleMR* package.[131] If results for a SNP were not available in the GWAS summary results then a proxy (r^2>0.6) was used. If more than one instrumental SNP was available then IVW regression was performed. If only one instrumental SNP was available then the Wald ratio was used to calculate the MR estimates.

2.2.4.2 MR-Egger

Sensitivity analysis were performed using MR-Egger regression, which is a pleiotropy-robust method used to assess the validity of a genetic instrument.[132] In MR analysis with multiple genetic variants, if any of the genetic variants used have a pleiotropic effect on the outcome then the causal estimates may be biased. MR-Egger regression is used to detect and correct for bias due to pleiotropy.

MR-Egger regression analyses were conducted in R (version 3.3.3) using the *mr_egger* function from the *MendelianRandomization* package (https://cran.r-project.org/package=MendelianRandomization). For example, to conduct MR-Egger regression with BMI as the exposure trait and a metabolite as the outcome trait, where "G_bmi$coef" and "G_bmi$se" are vectors of the SNP-BMI coefficients and standard errors and "G_metabolite$coef" and "G_metabolite$se" are vectors of the SNP-metabolite coefficients and standard errors respectively, the following code could be used

```
MR.input.object <- mr_input(bx=G_bmi$coef, bxse=G_bmi$se,
                        by=G_metabolite$coef, byse=G_metabolite$se)

mr_egger(MR.input.object)
```

2.2.5. Mediation

A variable is a mediator of the causal relationship between an exposure and an outcome if it lies on the causal pathway between the exposure and the outcome.[17] Mediation can be partial or complete. Complete mediation occurs when the exposure can only affect the outcome through the mediator, and partial mediation occurs when other mechanisms exist through which the exposure can affect the outcome.

Mediation analyses were performed using the *mediate* function from the *mediation* package in R (version 3.3.3).[133] The mediated effect, the direct effect and the total effect were estimated. For example, to investigate whether a metabolite mediates the effect of diet on BMI, the following code could be used

```
med.fit <- lm(metabolite ~ diet + age + sex)

out.fit <- lm(bmi ~ metabolite + diet + age + sex)

med.out <- mediate(med.fit, out.fit, treat="diet",
                   mediator="metabolite", robustSE=TRUE, sims=1000)
```

CHAPTER 3. DIET GWAS

3.1. Introduction

3.1.1. Heritability of dietary intake

Strong evidence exists for a genetic basis for diet, though the magnitude of the role that genetics plays is less clear as effect estimates are heterogeneous. Family and twin studies estimate that genetic effects explain c. 20-40% of variation in energy and macronutrient intake.[46] However, genome-wide complex trait analyses (GCTA) have estimated that only c. 6-8% of variance in fat, protein and carbohydrate intake can be explained by common tag-SNPs.[134] Dietary behaviour is a complex trait, and hence few studies have identified genetic variants associated with diet. Heritability estimates from GCTA tend to be lower than those from twin studies since GCTA can only detect the additive effects of commons SNPs, but not gene-gene or gene-environment interactions, or other types of genetic variation such as copy number variation.[135]

There are many plausible ways in which the genome could affect dietary behaviour; these include appetite regulation, metabolism, satiety, absorption and mental health or behavioural traits. For example, a twin study has estimated the genetic heritability of satiety responsiveness to be 63%;[136] and a study of satiety responsiveness and genetic predisposition to obesity found that satiety responsiveness mediated the association between the obesity genetic risk score and adiposity.[137]

3.1.2. Previous diet GWAS

GWAS is a commonly used method for identifying genetic variants associated with a trait. GWAS have been successful in identifying SNP-trait associations for a wide range of complex traits, including traits such as educational attainment for which no replicable genetic associations had previously been identified.[138,139] An association between a SNP and a trait does not imply that the SNP directly influences the trait through a biological mechanism, but that the SNP is likely to be in linkage disequilibrium (LD) with a causal variant.[119]

So far, GWAS of energy and macronutrient intake have identified two replicable genetic associations, which were discovered in two concurrent GWAS of dietary macronutrient intake published in early 2013.[134,140] Tanaka et al. conducted a genome-wide meta-analysis of macronutrient intake in 37,537 participants from the CHARGE Consortium and attempted to replicate their findings in 33,533 participants from the DietGen Consortium;[140] and Chu et al. undertook a similar genome-wide meta-analysis in the DietGen Consortium and attempted to replicate their findings in the CHARGE Consortium.[134] Despite the moderately large sample sizes used in these GWAS, neither study observed a genome-wide significant association in their discovery analysis that replicated in the other study. Instead, the studies also took forward their top sub-genome-wide significant SNPs for replication and each succeeded in replicating one sub-genome-wide significant SNP (rs838145 and rs838133; both on 19q13.33) in the parallel study. When the results from two studies were meta-analysed, both SNPs reached genome-wide significance.

Two separate smaller GWAS were unable to detect any replicable genome-wide significant associations.[141,142] A GWAS of fat intake in 598 adolescents from a Canadian study did not observe any genome-wide significant associations; the smallest p-value observed was for rs2281617 in *OPRM1* ($p=5.2\times10^{-6}$).[141] A GWAS of confectionery intake in Japanese adults observed two genome-wide significant associations (rs2839525 and rs1147522) in the discovery phase (N=939 adults) however these associations did not hold in the replication phase (N=4,491 adults).[142]

The above GWASs of macronutrient intake collected diet data using either FFQs or a 24-food recall; the data was then summarised by estimating the proportions of total energy intake derived from each macronutrient studied.[134,140,141] Tanaka et al. observed genome-wide significant associations for rs838145 with carbohydrate and fat intake;[140] Chu et al. observed a genome-wide significant association between rs838133 and protein intake.[134] Since these macronutrients are measured as proportions of total energy intake they are not independent of each other, and hence it is not surprising that nearby genetic effects were identified for different macronutrients. Both GWAS identified *FGF21* as the top candidate gene in this region. *FGF21* encodes a hormone

involved in glucose and lipid metabolism.[140] Following these findings by Tanaka et al. and Chu et al., Soberg et al. investigated the relationship between *FGF21* and sweet food intake in the Danish Inter99 cohort.[143] Soberg et al. categorised sweet foods as either "candy" or "cake" and observed an association between rs838133 and candy intake but not cake intake.

The CHARGE and DietGen consortia GWASs performed genotyping using Illumina- or Affymetrix arrays and imputed to ~2.6 million SNPs (HapMap release 21 or 22/NCBI build 35 or 36). Haghighi et al. performed genotyping using Illumina Human610-Quad BeadChip (~570,000 SNPs), but did not impute SNPs.[141]

3.1.3. Challenges of diet GWAS and strengths of UK Biobank

Several things influence the ability of a complex trait GWAS to identify SNP-trait associations, including experimental sample size, the joint distribution of SNP effect size and allele frequency, and trait measurement error.[119] The GWAS in the CHARGE and DietGen consortia was conducted using a moderately large sample size, a validated method of macronutrient intake estimation, and a panel of ~ 2.6 million SNPs. Despite these assets, only two genome-wide significant associations were identified. This suggests that macronutrient intake is a highly heterogeneous trait influenced by many small genetic effects. Hence a larger sample size, a larger panel of SNPs, or a more precise method of dietary intake measurement will be required to identify further genetic variants associated with macronutrient intake.

UK Biobank has both dietary and genetic data available for ~144,000 people, which is more than twice the size of the previous largest macronutrient GWAS sample size (~71,000 participants from the combined samples from the CHARGE and DietGen consortia). UK Biobank participants were genotyped using the UK BiLEVE array or the UK Biobank axiom array, which contain ~800,000 markers. Genotype data in UK Biobank was imputed to the HRC reference panel (~40 million SNPs, of which ~11 million SNPs remain after filtering).[114] Dietary intake in UK Biobank was measured using a web-based 24-hour recall questionnaire administered at multiple timepoints, allowing for summary data to be averaged across timepoints and hence reducing the impact of daily variation

on the data. In light of previous macronutrient GWAS findings from other studies, the advantages in UK Biobank of a larger sample size, repeated diet measurements, and more genetic markers make it reasonable to expect to identify some diet-SNP associations.

3.1.4. Motivation for a diet GWAS

Dietary behaviour is known to be associated with a wide range of adverse health outcomes such as diabetes and atherosclerosis. Identifying genetic variants associated with dietary behaviour may lead to a better understanding how these adverse health outcomes may be prevented or treated – this is the primary motivation for these analyses. A second major motivation for conducting a diet GWAS in UK Biobank is to identify genetic variants that could be used to generate a robust genetic instrument for dietary intake. This genetic instrument could be used in MR to estimate bidirectional causal effects of diet on BMI.

3.2. Methods

UK Biobank is a population-based prospective cohort consisting of ~500,000 participants aged between 40 and 69 years who were recruited from across the UK (**2.1.2**).[144]

UK Biobank participants were genotyped using the UK BiLEVE array or the UK Biobank axiom array,[113] and their genotype data was imputed to the HRC reference panel.[114] Quality control filtering of the data is described in Chapter 2 (**2.1.2.1**).

Dietary data in UK Biobank was collected using a 24-hour recall questionnaire (http://biobank.ctsu.ox.ac.uk/crystal/docs/DietWebQ.pdf) (**2.1.2.2**). Diet data from the assessment visit is available for the last ~70,000 participants recruited; all participants were invited, via email, to complete the questionnaire at four later occasions. This dietary data was used to estimate energy and macronutrient intake during each 24-hour recall period.

3.2.1. Diet GWAS in UK Biobank

GWAS were conducted to identify SNPs associated with energy and macronutrient intake. A GWAS involves testing the relationship between each SNP and the trait. An association is "genome-wide significant" if the p-value is less than 5×10^{-8}. Typically, when conducting a GWAS, discovery analyses are performed to identify genome-wide significant SNPs and replication analyses are then performed in a separate unrelated sample to assess whether these associations hold.

Diet data has been collected in UK Biobank at the assessment visit and from four later online assessments. For these genome-wide analyses of dietary intake the UK Biobank cohort is split into two groups. Participants were assigned to the first group (the "visit" group) if they had completed the diet questionnaire at the assessment visit (n≈71,000). Participants were assigned to the second group (the "online" group) if they had completed at least one online questionnaire but had not completed a diet questionnaire at the assessment visit (n≈140,000). The benefit of splitting the cohort in two and running two parallel analyses is that it allows for any genome-wide significant SNPs found in one group to be "replicated" in the other group. Dietary intake was assessed in each group using the same 24-hour dietary recall questionnaire, however the questionnaire was administered in different formats – in one group the questionnaire was completed during the assessment centre visit and in the other group the questionnaire was sent out by email and completed online.

Participants in the "online" group have diet data available from between one and four online questionnaires (~49,000 participants completed only one online questionnaire, ~36,500 completed two, ~32,500 completed three, ~22,000 completed four). To reduce the impact of day to day variation on estimated dietary intake, average dietary intake variables were generated for each participant using data from all online questionnaires completed by that participant, and these are the dietary intake variables used in the GWAS. In the "visit" group, dietary intake was estimated from the diet questionnaire completed at the assessment visit.

GWAS were performed using a pipeline developed within the MRC-IEU. Quality Control filtering of the UK Biobank data was conducted by R.Mitchell, G.Hemani, T.Dudding, L.Paternoster as described in the published protocol (doi:10.5523/bris.3074krb6t2frj29yh2b03x3wxj).[115] The MRC IEU UK Biobank GWAS pipeline was developed by B.Elsworth, R.Mitchell, C.Raistrick, L.Paternoster, G.Hemani, T.Gaunt (doi: 10.5523/bris.2fahpksont1zi26xosyamqo8rr).[120]

More details about the pipeline are found in the methods chapter (**2.1.2.1** and **2.2.2.1**). In brief, GWAS were conducted using linear regression implemented in PLINK v2.00. The sample was restricted to individuals of white British ancestry. SNPs were restricted to autosomal SNPs within the HRC site list.

GWAS were performed for each of the following dietary intake estimates in UK Biobank: total energy (kJ/day); protein (g/day); total fat (g/day); carbohydrate (g/day); saturated fat (g/day); polyunsaturated fat (g/day); total sugars (g/day); and Englyst dietary fibre (g/day).[118]

Two models were fitted for each diet variable. The first model was adjusted for genotype array, sex and the first 10 genetic PCs. The second model was adjusted for BMI and the covariates from the first model. These models explore two slightly different questions about the relationship between genetic variation and dietary intake. The model without adjustment for BMI identifies any genetic association with diet, regardless of whether this relationship is mediated through BMI. The model with adjustment for BMI accounts for genetic effects mediated through BMI, and hence identifies genetic signals that may have a more direct effect on diet.

The BMI measure used in these GWAS is BMI measured at the assessment visit since this is the only BMI measure available for all participants. Hence, in the “visit” group BMI is measured at the same timepoint as diet, but in the “online” group BMI was measured at an earlier timepoint than diet.

3.2.2. LD score regression

LD score regression is a technique that uses GWAS summary data to estimate heritability and cross-trait genetic correlation.[145,146] LD score regression was performed to estimate the SNP heritability and pairwise cross-trait genetic correlation of the dietary traits using summary statistics from the energy and macronutrient GWASs performed in the "online" group. Cross-trait genetic correlation was estimated to assess the overlap in genetic variation driving different traits, and to compare the genetic correlation estimates with phenotypic correlation.

Analyses were run in R (version 3.0.2) using scripts based on those available at Github (https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation).[146] SNPs were restricted to HapMap 3 SNPs.

Cross-trait genetic correlation between the different diet traits and between the diet traits and BMI was estimated using the LD Hub online platform (http://ldsc.broadinstitute.org/).[147] The BMI GWAS summary data used are from Speliotes et al. (2010).[148] SNPs were restricted to HapMap 3 SNPs without the MHC region.

3.3. Results

3.3.1. Diet data in UK Biobank

Table 3 shows correlations between the "visit" and "online" diet measures for participants in the "visit" group.

Table 3 - Correlation between visit group and online group diet measures.

Diet measure	Correlation
Energy	0.40
Protein	0.31
Fat	0.36
Carbohydrates	0.43
Saturated fat	0.38
Polyunsaturated fat	0.25
Total sugars	0.46
Fibre	0.43

3.3.2. Diet GWAS in UK Biobank

GWAS were performed for eight dietary intake variables generated from 24-hour dietary recall questionnaires. Two models were fitted: the first model was adjusted for sex, genotype array and the first 10 genetic PCs, and the second model was adjusted for all the covariates in model 1 and BMI. Analyses were done in two separate samples: in the "online" group dietary intake was measured by averaging data from up to four online questionnaires; in the "visit" group daily dietary intake was measured using data from the questionnaire completed at the assessment centre visit. SNPs with MAF<1% were excluded from results.

In the "visit" group no genome-wide significant associations were observed. In the "online" group 43 genome-wide significant associations were observed, and at least one genome-wide significant association was observed for each diet phenotype (**Table 4**). Since many of these genome-wide significant hits were in high LD, only the top SNP from each LD block was selected for replication. SNPs were determined to be in the same LD block if LD between the SNPs was $R^2 > 0.8$ (in British population group from 1000 Genomes Project; LD estimate from LDlink https://analysistools.nci.nih.gov/LDlink/).

Manhattan and QQ plots for each of the diet traits in the "online" group, without adjustment for BMI, can be found in **Figure 3**.

Replication analyses for the following diet-SNP associations identified in the "online" group were performed in the "visit" group: rs7957145 with energy intake; rs838133 and rs13447258 with protein intake; rs13111413 and rs8097589 with carbohydrate intake; rs72828557 with fat intake; rs72828557 with saturated fat intake; rs516246 with polyunsaturated fat intake; rs200553669 with fibre intake; and rs2842189 and rs13111413 with total sugars intake (**Table 5**). The only diet-SNP association to survive multiple testing (Bonferroni-adjusted p<0.05/11 since analysing 11 diet-SNP associations) and replicate was rs838133 (maps to *FGF21*, protein coding gene) with protein intake. Four other diet-SNP associations also produced low p-values: rs72828557 (maps to *LOC107986574*, uncharacterised gene) with fat intake and saturated fat intake; rs516246 (maps to *FUT2*, protein coding gene) with polyunsaturated fat intake; and

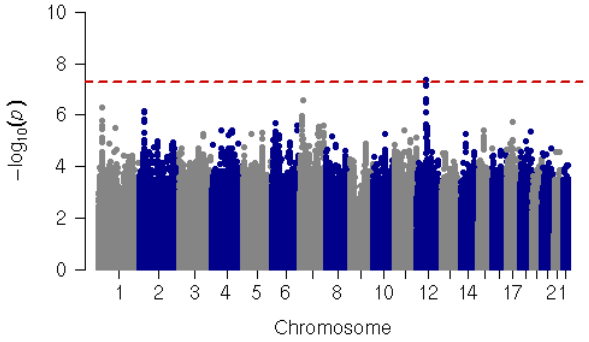
rs2842189 (maps to *PTPRF*, protein coding gene) with total sugars intake. The LD between SNPs rs838133 and rs516246 is $R^2 = 0.364$ (in British population group from 1000 Genomes Project; LD estimate from LDlink https://analysistools.nci.nih.gov/LDlink/). Information on the genes to which these SNPs map can be found in **Table 6**.

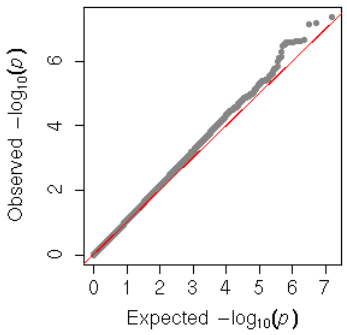
A fixed effect, inverse-variance weighted meta-analysis of the "online" and "visit" group diet-SNP results was performed for each of the SNPs taken forward for replication (**Table 5**).

The forest plot in **Figure 4** shows the results from this meta-analysis, along with results from the "online" and "visit" groups separately. Several of the diet-SNP effect estimates are highly concordant between the "online" and "visit" groups: protein and rs838133; fat and rs72828557; saturated fat and rs72828557; polyunsaturated fat and rs516246; total sugars and rs2842189. Some diet-SNP effect estimates are discordant: energy and rs7957145; carbohydrates and rs13111413; total sugars and rs13111413; fibre and rs200553669. BMI adjustment has a negligible influence on the effect sizes.

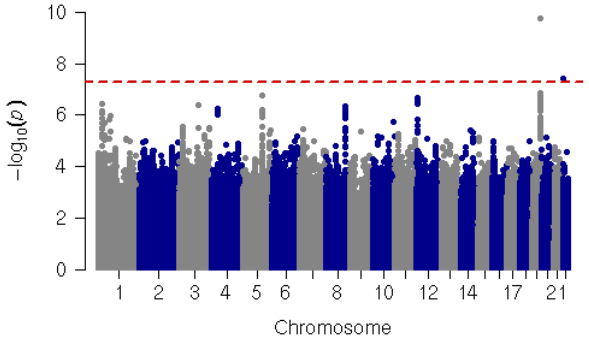
Figure 3 – Manhattan plots and QQ plots of results from dietary intake GWAS in the "online" group, without adjustment for BMI.

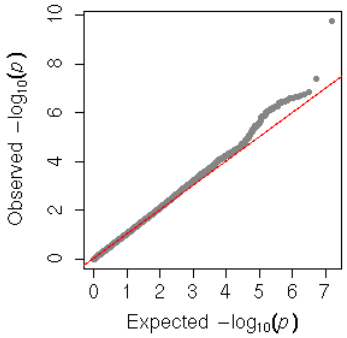
Energy intake:



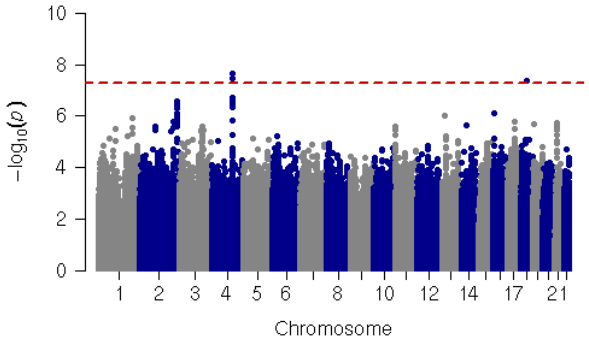


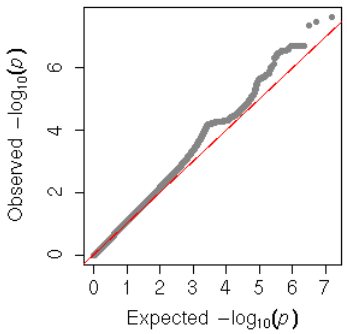
Protein intake:



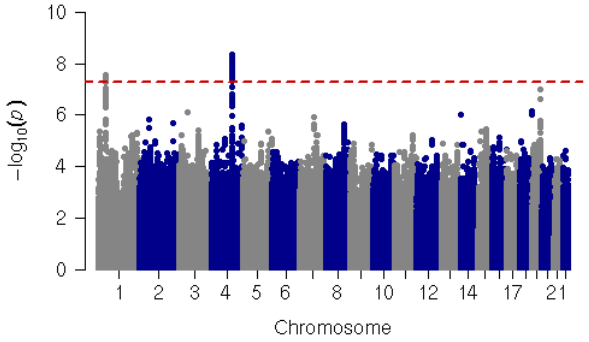


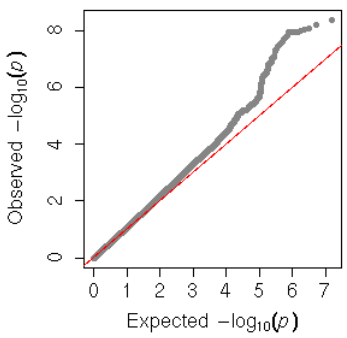
Carbohydrate intake:



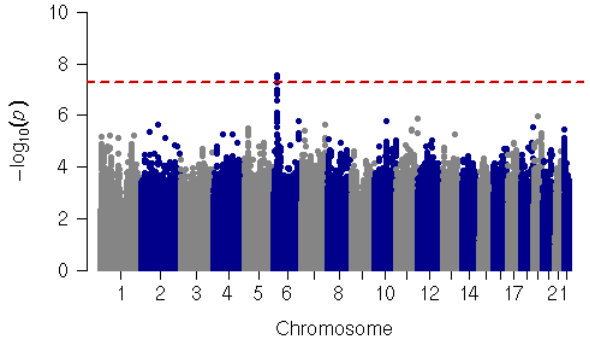


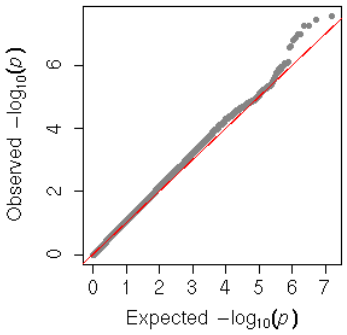
Total sugars intake:



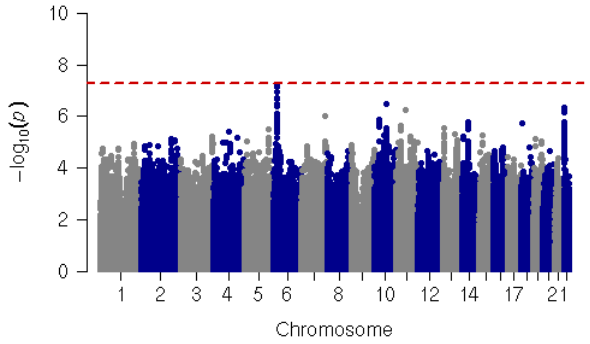


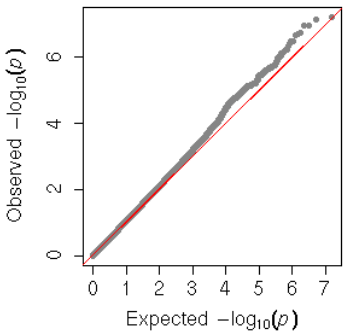
Fat intake:



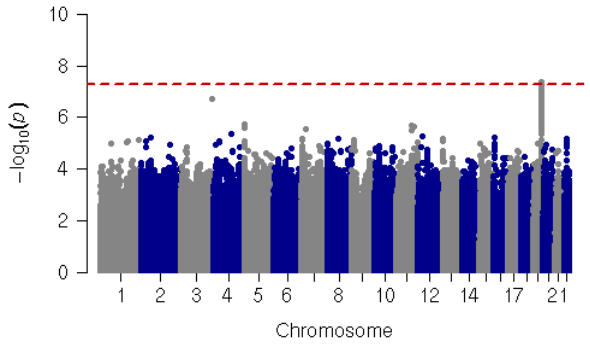


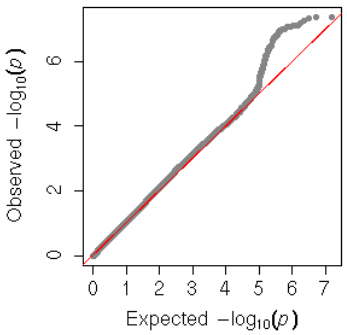
Saturated fat intake:



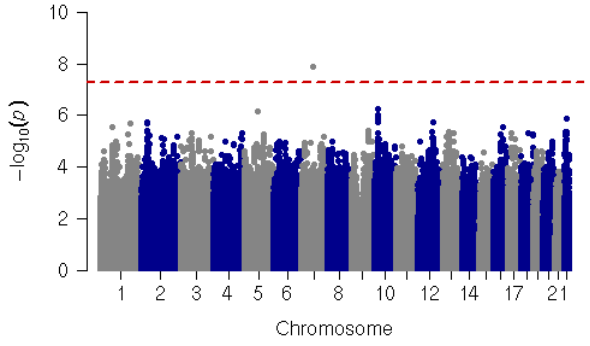


Polyunsaturated fat intake:





Fibre intake:



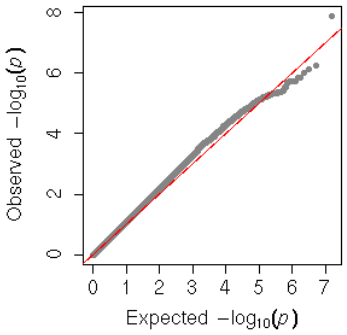


Figure 4 – Forest plot of the top diet-SNP associations.

Effect sizes are the macronutrient (g/day) increase per copy of the effect allele. Energy results are not included in this plot since the scale is different – (kJ/day) not (g/day).

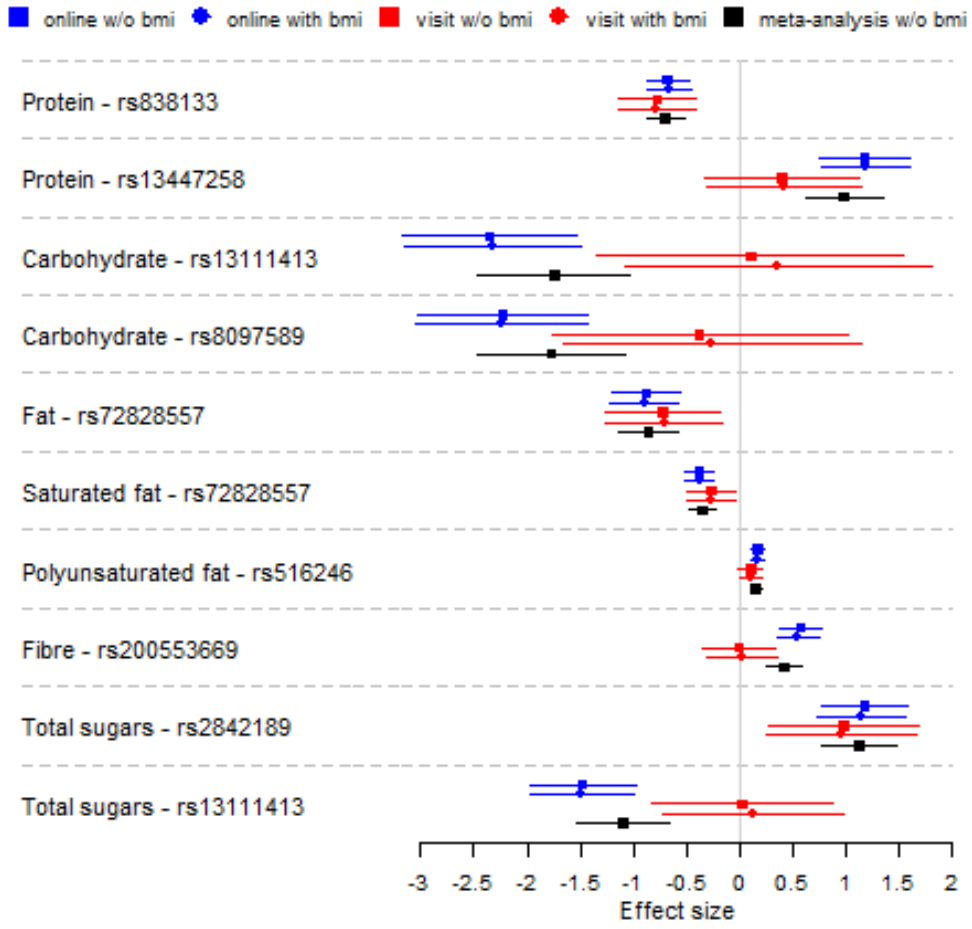


Table 4 – GWAS results with $p<5\times10^{-8}$ in the "online" group.

Chr, chromosome; BP, base position; MAF, minor allele frequency; MA, minor allele; Info, imputation information score; EA, effect allele. In models without adjustment for BMI, N=97,464-97,535. In models with adjustment for BMI, N=96,402-96,470.

								Not adjusting for BMI			Adjusting for BMI		
Trait	**SNP**	**Chr**	**BP**	**MAF**	**MA**	**Info**	**EA**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
Energy (kJ/d)	rs7957145	12	57302120	0.11	T	1.00	C	91.87	59.02, 124.73	4.25×10^{-8}	94.21	61.19, 127.23	2.24×10^{-8}
	rs34285886	12	57302520	0.11	G	1.00	A	89.97	57.36, 122.59	6.43×10^{-8}	92.10	59.32, 124.87	3.64×10^{-8}
	rs1391708	12	57305580	0.11	G	1.00	A	89.56	56.93, 122.19	7.48×10^{-8}	91.71	58.92, 124.50	4.21×10^{-8}
Protein (g/d)	rs838133	19	49259529	0.45	A	0.94	A	-0.68	-0.88, -0.47	1.68×10^{-10}	-0.66	-0.87, -0.46	4.00×10^{-10}
	rs13447258	22	19494074	0.07	A	0.91	G	1.17	0.75, 1.59	3.91×10^{-8}	1.18	0.76, 1.60	3.39×10^{-8}
Carb. (g/d)	rs13111413	4	129862368	0.21	T	0.99	C	-2.35	-3.17, -1.53	2.30×10^{-8}	-2.32	-3.15, -1.49	3.77×10^{-8}
	rs11736731	4	129863332	0.21	C	0.99	T	-2.32	-3.15, -1.50	3.30×10^{-8}	-2.29	-3.12, -1.47	5.47×10^{-8}
	rs8097589	18	40991505	0.23	A	0.98	G	-2.23	-3.03, -1.43	4.34×10^{-8}	-2.24	-3.05, -1.44	4.15×10^{-8}
Fat (g/d)	rs72828545	6	19108497	0.18	G	0.99	A	-0.88	-1.20, -0.56	5.52×10^{-8}	-0.90	-1.22, -0.58	3.58×10^{-8}
	rs112172280	6	19118597	0.18	C	0.99	T	-0.88	-1.20, -0.56	5.35×10^{-8}	-0.90	-1.22, -0.58	3.49×10^{-8}
	rs72828557	6	19128366	0.18	T	0.99	G	-0.90	-1.22, -0.59	2.69×10^{-8}	-0.92	-1.24, -0.60	1.73×10^{-8}
	rs72828558	6	19128832	0.18	C	0.99	T	-0.90	-1.21, -0.58	3.51×10^{-8}	-0.91	-1.23, -0.59	2.28×10^{-8}
Sat. fat (g/d)	rs72828557	6	19128366	0.18	T	0.99	G	-0.38	-0.52, -0.24	5.99×10^{-8}	-0.38	-0.52, -0.25	5.00×10^{-8}
Polyun. fat (g/d)	rs679574	19	49206108	0.49	C	1.00	C	0.17	0.11, 0.23	4.49×10^{-8}	0.16	0.10, 0.22	6.79×10^{-8}
	rs516316	19	49206145	0.49	G	1.00	G	0.17	0.11, 0.23	4.52×10^{-8}	0.16	0.10, 0.22	6.85×10^{-8}
	rs516246	19	49206172	0.49	C	1.00	C	0.17	0.11, 0.23	4.25×10^{-8}	0.16	0.11, 0.22	6.54×10^{-8}
Fibre (g/d)	rs200553669	7	73872749	0.04	G	0.48	A	0.57	0.37, 0.76	1.25×10^{-8}	0.55	0.35, 0.75	3.66×10^{-8}
Total sugars (g/d)	rs2152113	1	43983569	0.38	T	1.00	T	1.14	0.73, 1.55	4.54×10^{-8}	1.12	0.71, 1.53	8.97×10^{-8}
	rs11577403	1	43989773	0.38	A	0.99	G	-1.14	-1.55, -0.73	4.50×10^{-8}	-1.12	-1.53, -0.71	8.68×10^{-8}
	rs2842189	1	44007648	0.38	T	1.00	T	1.17	0.76, 1.58	2.64×10^{-8}	1.15	0.73, 1.56	5.22×10^{-8}
	rs2782640	1	44009033	0.38	C	1.00	C	1.16	0.75, 1.57	3.37×10^{-8}	1.13	0.72, 1.55	7.10×10^{-8}
	rs951740	1	44011737	0.38	G	1.00	G	1.15	0.74, 1.56	3.72×10^{-8}	1.13	0.72, 1.54	7.73×10^{-8}
	rs2782641	1	44013355	0.39	G	0.99	G	1.16	0.75, 1.57	2.87×10^{-8}	1.13	0.72, 1.54	7.66×10^{-8}
	rs13114904	4	129785512	0.20	A	0.99	G	-1.47	-1.97, -0.96	1.18×10^{-8}	-1.45	-1.96, -0.95	1.75×10^{-8}
	rs13125643	4	129789981	0.20	C	0.99	T	-1.42	-1.92, -0.91	3.28×10^{-8}	-1.40	-1.91, -0.9	5.06×10^{-8}
	rs11933240	4	129793559	0.20	G	0.99	A	-1.41	-1.91, -0.91	3.60×10^{-8}	-1.40	-1.9, -0.89	5.55×10^{-8}
	rs11730068	4	129796469	0.20	A	1.00	C	-1.40	-1.9, -0.9	4.80×10^{-8}	-1.38	-1.88, -0.88	7.49×10^{-8}
	rs13110952	4	129797144	0.20	T	0.99	C	-1.47	-1.97, -0.96	1.14×10^{-8}	-1.45	-1.96, -0.95	1.71×10^{-8}
	rs13139971	4	129801118	0.21	G	1.00	A	-1.40	-1.89, -0.91	2.57×10^{-8}	-1.38	-1.87, -0.89	4.16×10^{-8}
	rs11940298	4	129803970	0.21	G	1.00	A	-1.44	-1.94, -0.95	9.24×10^{-9}	-1.43	-1.92, -0.93	1.51×10^{-8}
	rs11945441	4	129811178	0.21	G	1.00	C	-1.45	-1.94, -0.95	8.61×10^{-9}	-1.43	-1.93, -0.94	1.34×10^{-8}
	rs13146706	4	129812702	0.20	G	1.00	A	-1.47	-1.97, -0.96	1.11×10^{-8}	-1.46	-1.96, -0.96	1.43×10^{-8}
	rs10518538	4	129813710	0.20	A	1.00	T	-1.47	-1.97, -0.96	1.09×10^{-8}	-1.46	-1.96, -0.96	1.41×10^{-8}
	rs13135764	4	129813937	0.20	G	1.00	T	-1.47	-1.97, -0.96	1.10×10^{-8}	-1.46	-1.96, -0.96	1.41×10^{-8}
	rs11098991	4	129814440	0.20	G	1.00	A	-1.47	-1.97, -0.96	1.09×10^{-8}	-1.46	-1.97, -0.96	1.40×10^{-8}
	rs10857133	4	129828601	0.20	C	1.00	G	-1.47	-1.97, -0.97	9.97×10^{-9}	-1.47	-1.97, -0.96	1.25×10^{-8}
	rs11735343	4	129837470	0.20	C	1.00	T	-1.45	-1.95, -0.94	1.65×10^{-8}	-1.44	-1.95, -0.94	2.02×10^{-8}
	rs11722253	4	129840234	0.20	C	1.00	T	-1.45	-1.95, -0.95	1.53×10^{-8}	-1.45	-1.95, -0.94	1.88×10^{-8}
	rs11735359	4	129846496	0.20	A	1.00	G	-1.44	-1.95, -0.94	1.78×10^{-8}	-1.44	-1.94, -0.94	2.19×10^{-8}
	rs13133798	4	129858318	0.20	T	1.00	A	-1.44	-1.94, -0.94	2.00×10^{-8}	-1.43	-1.94, -0.93	2.56×10^{-8}
	rs13149221	4	129860709	0.20	T	0.99	C	-1.43	-1.94, -0.93	2.70×10^{-8}	-1.42	-1.93, -0.92	3.61×10^{-8}
	rs13111413	4	129862368	0.21	T	0.99	C	-1.48	-1.97, -0.98	4.19×10^{-9}	-1.49	-1.98, -0.99	3.51×10^{-9}
	rs11736731	4	129863332	0.21	C	0.99	T	-1.46	-1.95, -0.97	6.31×10^{-9}	-1.47	-1.96, -0.98	5.39×10^{-9}

Table 5 – Replication of top associations from the "online" group in the "visit" group; meta-analysis of results from both groups.

MA, minor allele; EA, effect allele. In models without adjustment for BMI, N = 47,134 – 47,197. In models with adjustment for BMI, N = 46,618 – 46,679.

			"Visit" group						Meta-analysis of "online" and "visit" groups		
			Without adjusting for BMI			With adjusting for BMI			Without adjusting for BMI		
Trait	**SNP**	**EA**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
Energy (kJ/d)	rs7957145	T	-2.99	-60.71, 54.73	0.919	-4.78	-62.73, 53.17	0.872	68.66	40.10, 97.21	2.45×10^{-6}
Protein (g/d)	rs838133	G	-0.78	-1.14, -0.41	2.96×10^{-5}	-0.78	-1.15, -0.41	3.04×10^{-5}	-0.70	-0.88, -0.52	2.58×10^{-14}
	rs13447258	A	0.39	-0.34, 1.12	0.299	0.42	-0.32, 1.15	0.263	0.98	0.62, 1.34	1.23×10^{-7}
Carbs (g/d)	rs13111413	T	0.10	-1.34, 1.53	0.895	0.36	-1.08, 1.80	0.626	-1.74	-2.46, -1.03	1.77×10^{-6}
	rs8097589	A	-0.38	-1.77, 1.02	0.598	-0.26	-1.66, 1.14	0.715	-1.77	-2.47, -1.08	5.35×10^{-7}
Fat (g/d)	rs72828557	T	-0.72	-1.27, -0.18	0.010	-0.71	-1.26, -0.17	0.011	-0.86	-1.13, -0.58	9.92×10^{-10}
Sat. fat (g/d)	rs72828557	T	-0.27	-0.50, -0.03	0.025	-0.27	-0.50, -0.03	0.026	-0.35	-0.47, -0.23	6.38×10^{-9}
Polyun. fat (g/d)	rs516246	T	0.10	-0.01, 0.20	0.064	0.11	0.00, 0.21	0.047	0.15	0.10, 0.20	1.36×10^{-8}
Fibre (g/d)	rs200553669	G	-0.01	-0.35, 0.33	0.962	0.02	-0.32, 0.36	0.915	0.42	0.25, 0.59	9.05×10^{-7}
Total sug. (g/d)	rs2842189	C	0.98	0.27, 1.69	0.007	0.95	0.24, 1.67	0.009	1.12	0.76, 1.48	6.79×10^{-10}
	rs13111413	T	0.02	-0.83, 0.87	0.967	0.13	-0.72, 0.98	0.767	-1.10	-1.53, -0.67	4.13×10^{-7}

Table 6 – Gene information for the diet-SNP associations that replicated.

The dbSNP database (https://www.ncbi.nlm.nih.gov/snp/) was used to identify the nearest gene to each SNP. The GWAS catalog (https://www.ebi.ac.uk/gwas/) was used to find published associations between the SNP/gene and any traits (traits in bold below if also appeared when searching for that particular SNP). GeneCards (http://www.genecards.org/) was used to look-up the molecular function of each gene.

SNP	Associated macronutrient(s)	Mapped gene	GWAS catalog trait	Molecular function (GeneCards)
rs838133	protein	*FGF21*	bipolar disorder,[149] **dietary macronutrient intake**,[134,140] **homocysteine levels**,[150] resting metabolic rate,[151] retinal vascular caliber[152]	stimulates glucose uptake in differentiated adipocytes[153,154]
rs72828557	fat, saturated fat	*LOC107986574*	none	-
rs516246	polyunsaturated fat	*FUT2*	bipolar disorder,[149] blood metabolite levels,[155] childhood ear infection,[156] cholesterol,[157] **Crohn's disease**,[158-161] diarrhoeal disease at age 1,[162] dietary macronutrient intake,[134] elevated serum carcinoembryonic antigen levels,[163] folate pathway vitamin levels,[164,165] homocysteine levels,[150] **inflammatory bowel disease**,[161] **liver enzyme levels**,[166] lung adenocarcinoma,[167] metabolic traits,[168] **obesity-related traits**,[169] paediatric autoimmune diseases,[170] primary sclerosing cholangitis,[171] psoriasis,[172,173] resting metabolic rate,[151] retinal vascular caliber,[152] serum lipase activity,[174] tumour biomarkers,[175] urinary metabolites,[176] vitamin B levels in ischemic stroke,[177] vitamin B12 levels[178-180]	influences secretor status and intestinal microbiota composition,[181,182] interacts with Crohn's Disease to influence colonic mucosa-associated microbiota[183]
rs2842189	total sugars	*PTPRF*	amyotrophic lateral sclerosis (age of onset),[184] autism spectrum disorder or schizophrenia,[185] educational attainment,[186] schizophrenia[187]	involved in cell signalling

3.3.3. Follow up of diet-SNP associations from the literature

SNP look-ups of these GWAS results were done for the following SNP-diet associations from the literature: rs838145 with carbohydrate intake and fat intake (Tanaka et al., proportions of energy derived from carbohydrate and fat intake); rs838133 with protein intake (Chu et al., proportion of energy derived from protein intake); and rs2839525 and rs1147522 with total sugar intake (Wakai et al., confectionary intake frequency) (**Table 7**).[134,140,142]

The only association to replicate in both the "online" and the "visit" groups was rs838133 with protein intake. The associations between rs838145 and carbohydrate and fat intake replicated in the "online" group but not the "visit" group, however the effect direction was consistent in both groups. The associations between rs1147522, rs2839525 and total sugars intake did not replicate.

In UK Biobank protein intake is defined as grams per day (g/d), whereas in the analyses by Chu et al. they quantified protein intake as percentage of total caloric intake from protein, so the effect sizes cannot be compared, however the direction of effect here is consistent with that reported by Chu et al.[134] Similarly, the UK Biobank effect sizes and the Tanaka et al. effect sizes cannot be compared, though they are directionally consistent.

3.3.4. Heritability and correlation

LD score regression was performed to estimate the heritability of the diet traits and explore the shared genetic architecture of gene variants between diet traits and with BMI (**Table 8**). The heritability estimates are low (~3-6%), though only a little lower than the GCTA heritability estimates from Chu et al. (~6-8%).[134]

The genetic correlation between diet traits is high (~25-96%), especially between energy and fat, and energy and carbohydrate. All genetic correlations between diet traits are positive. Genetic correlation is weakest between total sugars and protein, total sugars and polyunsaturated fat, and between fibre and fat. Saturated fat correlates more

strongly than polyunsaturated fat with all other diet traits except fibre. Overall, total sugars and fibre correlate least strongly with all the other diet traits.

Genetic correlations with BMI are weaker. Each of the diet traits is positively correlated with BMI, except for protein which is negatively correlated with BMI.

Table 9 shows the phenotypic correlation estimates between each of the diet traits (in the "online" group). When comparing the phenotypic correlations with the genetic correlations, in most cases the genetic correlations are stronger. This difference is greatest for saturated fat and polyunsaturated fat, where the genetic correlation between them is 0.76 and the phenotypic correlation between them is only 0.45.

Table 7 – Follow up of diet-SNP associations from the literature.

Direction of effect; effect allele; p-value	Literature	"Online" group (without adj. BMI)	"Online" group (with adj. BMI)	"Visit" group (without adj. BMI)	"Visit" group (with adj. BMI)
rs838145 with carbohydrate intake	+ve; G; $p=1.68\times10^{-8}$ (Tanaka et al.[140])	+ve; G; p=0.004	+ve; G; p=0.007	+ve; G; p=0.634	+ve; G; p=0.648
rs838145 with fat intake	-ve; G; $p=1.57\times10^{-9}$ (Tanaka et al.[140])	-ve; G; p=0.003	-ve; G; p=0.002	-ve; G; p=0.246	-ve; G; p=0.228
rs838133 with protein intake	-ve; A; 7.9×10^{-9} (Chu et al.[134])	-ve; A; $p=1.68\times10^{-10}$	-ve; A; $p=4.00\times10^{-10}$	-ve; A; $p=2.96\times10^{-5}$	-ve; A; $p=3.04\times10^{-5}$
rs1147522 with total sugars intake	+ve; T; 4.3×10^{-8} (Wakai et al.[142])	+ve; C; p=0.240	+ve; C; p=0.239	-ve; C; p=0.338	-ve; C; p=0.428
rs2839525 with total sugars intake	+ve; G; 5.5×10^{-9} (Wakai et al.[142])	-ve; T; p=0.841	-ve; T; p=0.861	+ve; T; p=0.393	+ve; T; p=0.135

Table 8 – Heritability and genetic correlation estimates from LD score regression

rg	Heritability	Energy	Protein	Carb.	Total fat	Sat. fat	Polyun. fat	Total sugars	Fibre	BMI
Energy	5.1%	1	0.67	0.86	0.91	0.90	0.78	0.66	0.46	0.14
Protein	3.3%	0.67	1	0.51	0.64	0.62	0.54	0.31	0.50	-0.15
Carb.	5.0%	0.86	0.51	1	0.68	0.68	0.58	0.90	0.63	0.13
Total fat	5.3%	0.91	0.64	0.68	1	0.96	0.89	0.38	0.31	0.18
Sat. fat	4.4%	0.90	0.62	0.68	0.96	1	0.76	0.39	0.25	0.09
Polyun. fat	3.9%	0.78	0.54	0.58	0.89	0.76	1	0.27	0.34	0.19
Total sugars	4.7%	0.66	0.31	0.90	0.38	0.39	0.27	1	0.60	0.11
Fibre	5.9%	0.46	0.50	0.63	0.31	0.25	0.34	0.60	1	0.02

Table 9 – Phenotypic correlation between diet traits

correlation	Energy	Protein	Carb.	Total fat	Sat fat	Polyun. fat	Total sugars	Fibre
Energy	1	0.71	0.83	0.84	0.77	0.63	0.62	0.47
Protein	0.71	1	0.50	0.60	0.53	0.44	0.36	0.38
Carb.	0.83	0.50	1	0.55	0.53	0.42	0.81	0.55
Total fat	0.84	0.60	0.55	1	0.89	0.75	0.35	0.32
Sat. fat	0.77	0.53	0.53	0.89	1	0.45	0.37	0.21
Polyun. fat	0.63	0.44	0.42	0.75	0.45	1	0.22	0.39
Total sugars	0.62	0.36	0.81	0.35	0.37	0.22	1	0.45
Fibre	0.47	0.38	0.55	0.32	0.21	0.39	0.45	1

3.4. Discussion

3.4.1. Main findings

This GWAS of dietary intake identified 11 genome-wide significant diet-SNP associations (43 including those in LD), of which 5 replicated, including rs516246 (in *FUT2*) with polyunsaturated fat intake, rs838133 (in *FGF21*) with protein intake and rs2842189 (in *PTPRF*) with total sugars intake.

The association between rs516246 and polyunsaturated fat intake is a novel GWAS finding. This SNP has been previously reported in GWAS of other traits, including Crohn's disease[159,161] and liver enzyme levels.[166] rs516246 is located in the *FUT2* gene, which influences secretor status and intestinal microbiota composition.[181,182] A study of 47 individuals (29 with Crohn's disease, 18 controls) observed several disease-by-genotype associations with intestinal microbiota.[183] Evidence for an effect of a dietary factor on risk of Crohn's disease is limited and conflicting,[188] whilst people with Crohn's disease are often advised that altering their diet may help their Crohn's symptoms. Further analysis will be needed to establish causality in these relationships (**Figure 5**).

An association between rs838133 (located in the *FGF21* gene) and protein intake has been previously reported in a GWAS meta-analysis by the DietGen and CHARGE consortia (beta=-0.11, $p=7.9 \times 10^{-9}$).[134] *FGF21* encodes a hormone involved in glucose and lipid metabolism. The association observed in this GWAS in UK Biobank is stronger (beta=-0.70, $p=2.58 \times 10^{-14}$) and directionally consistent with the finding by the DietGen and CHARGE consortia, thus further strengthening the evidence. As stated earlier, the effect sizes cannot be directly compared since in UK Biobank protein intake is defined as grams per day, whereas in the previous GWAS it was quantified as percentage of total caloric intake from protein, however, it is possible to do a more approximate comparison. A 2008-09 study of adults aged 19-64 years observed the following median protein intakes: protein (g/d) = 88.8 in men and 65.6 in women; protein (% food energy) = 16.8 in men and 17.1 in women.[189] So, a crude conversion of the UK Biobank beta from -0.70g/d to % energy intake would be beta = $-0.70 \div 0.5(88.8 + 65.6) \times 0.5(16.8 + 17.1) =$

-0.15 % energy intake, which is fairly comparable considering the crude method used here.

SNPs rs516246 and rs838133 are located ~53,000 base pairs apart on chromosome 19. The LD between the two SNPs is $R^2 = 0.364$ (in British population group from 1000 Genomes Project; LD estimate from LDlink https://analysistools.nci.nih.gov/LDlink/). Therefore, it is highly possible that both SNPs are tagging the same causal variant.

The association between rs2842189 (*PTPRF*) and total sugars intake is a novel GWAS finding. Previous GWAS which have observed associations with SNPs mapped to *PTPRF* include educational attainment,[186] schizophrenia[185,187] and age of onset amyotrophic lateral sclerosis (ALS).[184] A study of energy homeostasis in transgenic ALS mice observed an energetic deficit.[190] Compensating this deficit with a high-energy diet resulted in a 20% increase in mean survival, which suggests that energy intake may play a role in ALS.

Few studies have identified and replicated genome-wide significant diet-SNP associations. SNP look-ups in these GWAS results were performed for five SNP-diet associations from the literature (**Table 7**). The three diet-SNP associations which have replicated elsewhere also replicated in UK Biobank: rs838133 with protein intake, and rs838145 with carbohydrate and fat intake.[134,140] The associations between rs1147522, rs2839525 and total sugars intake have not been replicated previously, and did not replicate in UK Biobank.[142]

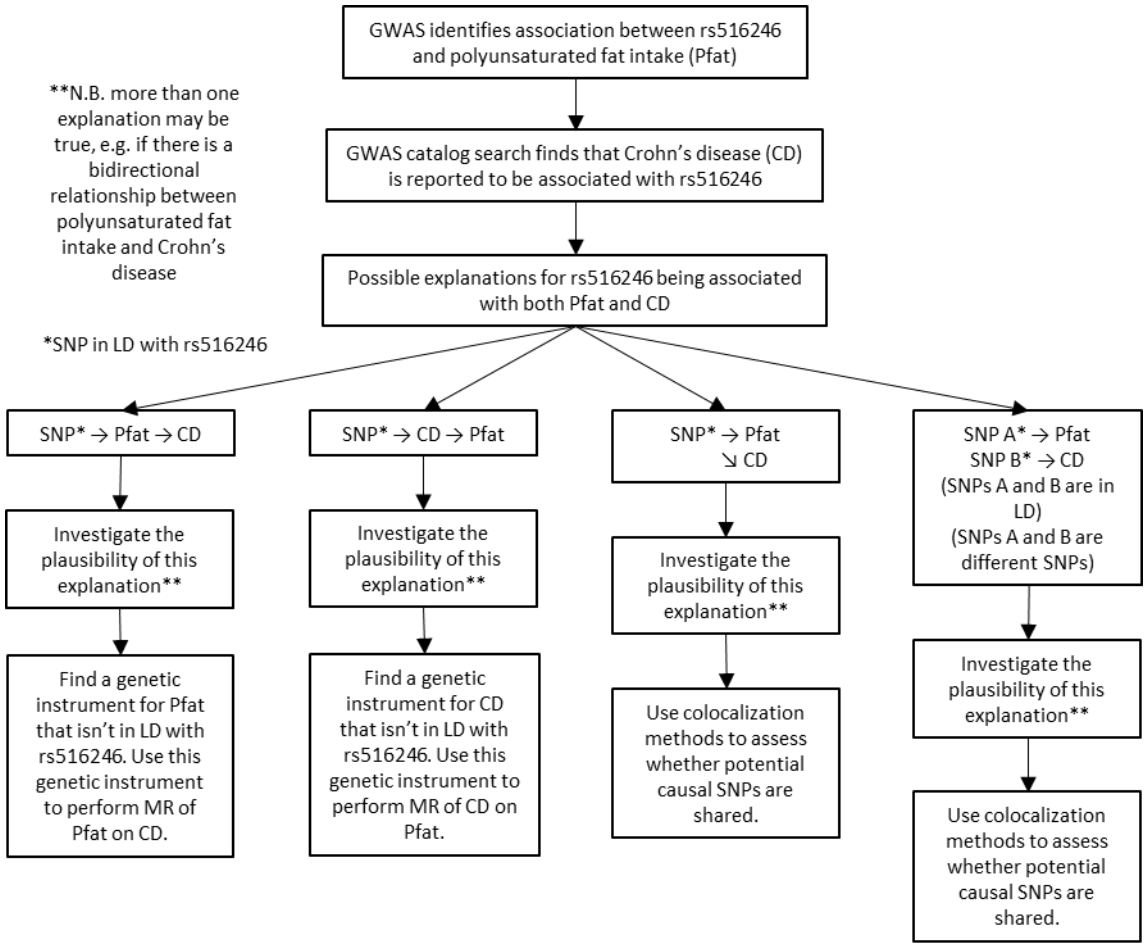
LD score regression was used to estimate pairwise genetic correlation between each of the diet traits. Both the pairwise genetic correlations and phenotypic correlations are high. This is to be expected given the complex composition of food and that, for example, saturated fat and polyunsaturated fat often co-occur, as do protein and fat, in same food stuff and hence genetic determinants of intake of one will correlate with the other.

When comparing the phenotypic correlations with the genetic correlations, in most cases the genetic correlations are stronger. A GWAS of 717 traits in UK Biobank that

estimated genetic and environmental correlations between pairs of traits found that for many of the pairs of traits the genetic and environmental correlation changes sign or the environmental correlation is stronger than the genetic correlation.[191] They concluded that the phenotypic covariance between many of the traits was many driven by environmental factors and not genetics.

Figure 5 – Flowchart of further analyses that could be conducted to explore why rs516246 is associated with both polyunsaturated fat intake and Crohn's disease.

The idea for this flowchart came from Richardson et al. (2017).[192] Colocalization is discussed in Fortune et al. (2015).[193]



3.4.2. Strengths and limitations

Identifying genetic variants associated with dietary behaviour is challenging since it is a complex phenotype that is hard to quantify. It is also difficult to differentiate between SNPs that are only associated with diet through their link with another phenotype such as BMI, and SNPs that have a more direct effect on diet.

The genome-wide significance threshold is usually defined as 5×10^{-8}. Since eight diet phenotypes are studied here, an alternative more conservative approach could be to apply a Bonferroni correction and set the threshold to be 6.25×10^{-9} ($=5 \times 10^{-8} / 8$), however this would be overly conservative due to the high correlation between the phenotypes. A less conservative approach would be to conduct PCA to calculate how many PCs are needed to account for ≥95% of variation in the eight diet phenotypes, and set the significance threshold as 5×10^{-8} divided by that number of PCs. This approach is used in chapter 5 on the metabolites since many are highly correlated.

Macronutrient intake was measured in g/day, rather than proportion of total energy intake. This allowed for consistency across the macronutrients since, whilst it is easy to comprehend fat intake as % energy intake, it is less common to measure of fibre intake in this way.

Quantifying dietary behaviour by estimated macronutrient intake has pros and cons. Macronutrient intake is an objective measure that many people are familiar with since food packaging often includes this information. However, the macronutrients included in this study do not fully describe a person's dietary choices, such as how often they consume processed food.

In this GWAS, dietary behaviour was recorded using 24-hour diet recall questionnaires. A limitation of such questionnaires is that dietary intake can vary considerably from day to day. This limitation was partially overcome in the "online" group administering the questionnaire multiple times and averaging the results. However, if a participant knows in advance that they need to complete a 24-hour recall questionnaire (this was the case for the UK Biobank online questionnaires, where participants were allowed at least

three days to complete the questionnaire), then this might influence their food choices in that 24-hour period so that they can record a smaller or "healthier" food intake.

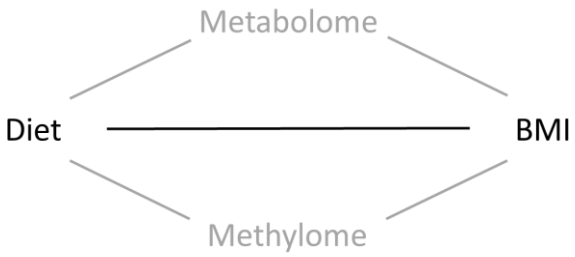
An alternative approach to measuring dietary behaviour could be to use metabolites as a biological proxy for macronutrient intake, for example amino acids as a proxy for protein intake. This is explored in Chapter 6. The advantage of using a biological proxy is that it does not rely on self-report, and therefore provides a more objective measure. It also represents the bioavailable proportion of the macronutrient rather than quantity eaten, and thus reflects absorption efficiency and other components of the digestive process over and above dietary intake.

3.4.3. Future directions

These findings provide motivation for future research. One area is to investigate the causal pathways that macronutrient genetic variants may play a part in, such as in **Figure 5**. Genetic variants robustly associated with specific macronutrients can be used in MR analysis to explore the long-term consequences of differences in intake of these factors.[194] For example, there is considerable interest in the influence of specific nutrients on risk of complex diseases such as cancer,[195] cardiovascular disease,[196,197] and mental health.[198] MR could help to strengthen causal inference with respect to the contribution that these dietary factors make to disease risk (or, in theory, disease progression[199]). This, in turn, could help to inform future interventions and help to overcome the often conflicting dietary advice that arises from more conventional observational epidemiology studies.

A major limitation of dietary studies is the blunt tools available for measuring dietary intake. Even the application of relatively advanced methods such as GWAS and MR is limited by reliance on the often biased self-report of dietary intake. Innovations that might improve this include the use of digital technologies, such as cameras and software to estimate food groups and portion size, or the use of more objective measures of dietary factors such as metabolites. These innovations are often time-consuming and/or expensive and hence studies have tended to rely on more traditional methods such as FFQs and diet diaries.

CHAPTER 4. DIET AND BMI



4.1. Introduction

4.1.1. Observational studies of diet and adiposity

Dietary behaviour and BMI are known to be strongly linked. However, the finer details of this relationship are less clear. Several studies have investigated the relationship between dietary behaviours and BMI or other measures of adiposity (e.g. body fat percentage and waist circumference). Most of these studies either observed associations between an "unhealthy" diet and increased adiposity or did not see any association, and a few studies found inverse associations. As discussed previously, dietary behaviour is complex and hence assessing the relationship between dietary behaviour and other phenotypes is often challenging.

Evidence from a systematic review of dietary energy density and body weight suggests that dietary energy density is positively associated with increased adiposity in children and adolescents.[18] A study of fast food consumption found that children who ate fast food during a typical day consumed more total energy and more energy per gram of food than those who did not; they also consumed less fibre and fewer fruits and non-starchy vegetables.[45] The authors hypothesised that fast food consumption may affect body weight through children replacing healthier food options with more energy-dense fast food.

A cross-sectional study of the relationship between macronutrient intake and body fat percentage in children aged 9 and 10 years old observed that adiposity was positively associated with percentage of energy derived from fat and negatively associated with percentage of energy derived from carbohydrate.[19]

Other studies observed inconsistent or weak evidence for a relationship between diet and adiposity measures in childhood and adolescence.[34,200,201] Some studies observed unexpected results, including inverse associations between consumption of 'unhealthy' snacks and adiposity.[29,200]

Several studies of diet and adiposity have been carried out in adults. A cross-sectional study of healthy older men found that obesity was associated with total energy intake and energy intake from fat.[22] A different study, also in males, found higher fibre intake levels to be associated with smaller increases in BMI and waist circumference but did not observe an association between fat or sugar intake and subsequent BMI or waist circumference.[28] A study in women observed associations between dietary patterns and weight change, including associations between high intakes of red and processed meats, refined grains, sweets and desserts and long-term weight gain.[23]

A systematic review of the role of dietary patterns in predicting weight change found evidence that a high intake of fibre and nuts predicts less weight gain, and a high meat intake predicts more weight gain.[24] A systematic review of dietary energy density and body weight found that low energy density dietary patterns aide weight maintenance and weight loss in adults.[18]

There is evidence that changing dietary behaviour may have a positive effect on obesity, including two longitudinal studies of adults which found that positive changes in eating behaviour (i.e. to a more "healthy" diet) were accompanied by a decrease in BMI or a smaller weight gain.[25,26]

In summary, these observational studies have identified links between increased adiposity and energy density, in particular fat intake. In contrast, a high fibre intake has been linked to a lower increase in adiposity.

4.1.2. Dietary interventions to combat obesity

Systematic reviews have been performed to study the effectiveness of weight loss interventions. A systematic review of RCTs investigating the effectiveness of dietary interventions of varying macronutrient distributions on weight loss in overweight or obese children and adolescents concluded that macronutrient distribution did not seem to affect weight loss (which contrasts with observational evidence above[19]), but instead dietary interventions should focus on reducing total energy intake.[202]

Another systematic review examined evidence for an association between sugar intake and body weight in children and adults.[203] Results from trials in adults comparing a trial arm in which participants were asked to reduce their sugar intake against a control arm provide evidence for a positive association between reduced sugar intake and weight loss. Intervention trials in which children were advised to reduce their sugar intake did not provide evidence for an association with BMI, however three of the five studies reported poor compliance.

4.1.3. Studies in UK Biobank

A study of the association between macronutrient intake and adiposity in UK Biobank participants found that fat was the largest contributor to overall energy intake, and that obesity was most strongly associated with total energy intake (from all macronutrients combined) rather than energy intake from any individual macronutrient.[35] When adjusting for total energy intake, fat intake was positively associated with obesity, whereas sugar intake was negatively associated with obesity. They used a categorical approach to perform their analyses: they calculated the average macronutrient intake for each BMI group (normal, overweight, etc.); and they calculated the average BMI for each quintile of macronutrient intake.

Another study in UK Biobank investigated whether macronutrient intake modifies the relationship between a BMI allele score and adiposity, and found evidence that the relationship is modified by total energy intake, total fat intake and, most strongly, by total saturated fat intake.[204] This relationship with fat and saturated fat intake remained even when models were adjusted for total energy intake, suggesting that the association is independent of total energy intake.

4.1.4. Studies in ALSPAC

Dietary data in ALSPAC has predominantly been collected in the form of food frequency questionnaires (FFQs) and diet diaries. These methods of data collection gather information on large numbers of different food items, so in order to identify dietary

patterns empirical methods, such as principal components analysis (PCA) or cluster analysis, may be used.[44]

An ALSPAC study of early life risk factors for childhood obesity looking at 25 putative risk factors including dietary behaviour did not find conclusive evidence for an association between dietary patterns (identified using PCA) at age 3 and obesity risk at age 7, but did observe a weak association between the junk food dietary pattern and risk of obesity.[205]

A study of the relationship between dietary behaviour and changes in body composition between the ages of 9 and 11 years in ALSPAC observed small associations between dietary pattern scores (from PCA) and changes in body composition.[20] A 'health aware' dietary pattern, in which there was a high intake of fruits and vegetables, high-fibre bread, cheese and fish and a low intake of fizzy drinks and processed foods including processed meat, was associated with a decrease in fat mass gain in girls between the ages of 9 and 11 years. A 'packed lunch' dietary pattern, characterised by high intakes of sandwiches and snacks, was associated with a decrease in fat mass gain in the girls and an increase in lean mass gain in the boys.

Another statistical method that has been used to extract diet patterns from ALSPAC diet data is reduced rank regression (RRR). RRR is used to identify patterns in a set of predictor variables that explain the maximum variation in a set of variables known as the "response" variables.[206] The response variables are hypothesised to be intermediate variables between the predictor variables and an outcome of interest and hence RRR makes use of prior knowledge, unlike PCA which is purely exploratory. For example, Ambrosini et al. defined their predictor variables to be food group intakes, their response variables to be dietary energy density and % energy from fat and fibre, and their outcome to be adiposity.[206] Studies implementing RRR in the ALSPAC participants across childhood and adolescence found that an energy-dense, low-fibre, high-fat diet was prospectively associated with greater fat mass and higher odds of excess adiposity.[206,207]

A study assessing the association between the consumption of fast food and BMI in ALSPAC found that teenagers who visited fast food outlets more frequently tended to be exposed to more unhealthy foods at home and have higher BMIs.[21] Teenagers who visited fast food outlets more frequently also consumed less fruit and vegetables, which is a further negative effect on their diet since in addition to consuming more saturated fat in unhealthy foods their consumption of important nutrients from fruit and vegetables is lower.

A study exploring the role of dietary intake in the relationship between *FTO* region BMI-associated SNPs and BMI in the ALSPAC children observed associations between *FTO* alleles and energy and fat intake, both before and after BMI adjustment.[208] Their results suggest that the association between *FTO* alleles and dietary behaviour is not solely due to the *FTO*-BMI association.

4.1.5. Motivation and objectives for these analyses

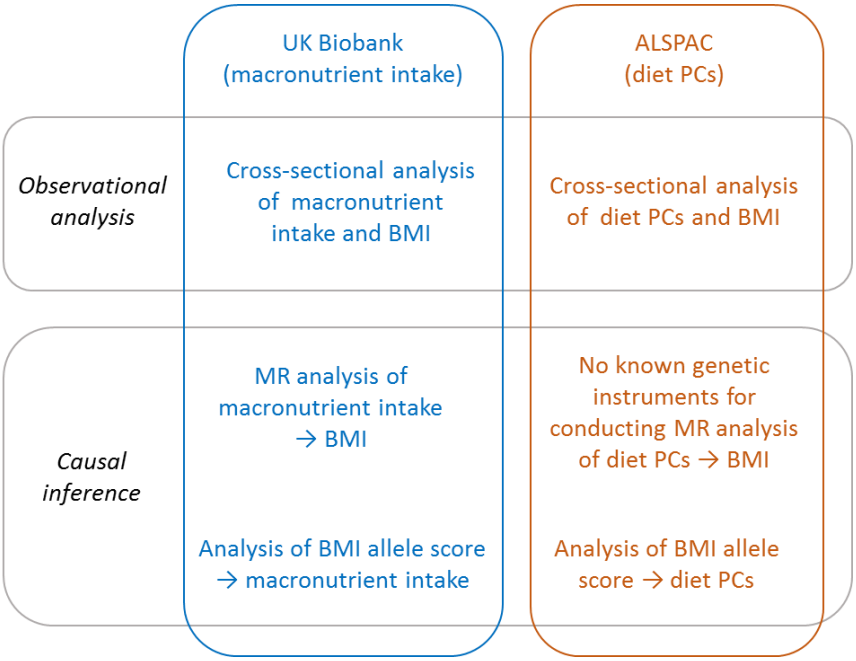
The relationship between dietary behaviour and adiposity is complex, and a better understanding of causality in this relationship is needed to improve public health policies aimed at tackling the obesity epidemic. The growing number of BMI-associated SNPs, and the discovery of some diet-associated SNPs (Chapter 3), may be applied to gain a clearer understanding this relationship.

Although Anderson et al. investigated the observational relationship between macronutrients and BMI, they did not explore causality.[35] They also only studied fat intake, and not saturated fat intake and polyunsaturated fat intake separately. They presented their results by BMI category (e.g. overweight) or quintile. Since the associations they observed are mainly linear, the analyses in this chapter will fit linear regression models (for all BMI categories together, not individually by category) for ease of interpretability and comparison with these previous findings.

Celis-Morales et al. found that the association between the GIANT BMI score and BMI is modified by fat intake. However, they did not explore the possibility that the GIANT BMI score may be partially exerting its effect on BMI through diet.[204]

The main objectives for the analyses in this chapter are to explore the relationship between various measures of dietary intake and BMI; and to investigate causality bidirectionally in any observed associations. Dietary intake is summarised in two main ways: in UK Biobank dietary intake is summarised as macronutrient intake and in ALSPAC dietary intake is summarized using PCA. A summary of the analyses conducted in this chapter is shown in **Figure 6**.

Figure 6 – Summary of analyses undertaken in this chapter.



4.2. Methods

4.2.1. Macronutrient intake and BMI analyses

Since both diet and BMI can fluctuate greatly over months or years, more stringent data cleaning was performed for these analyses (compared to the GWAS in the previous chapter).

A participant's data from a particular questionnaire was excluded if:

- any of their energy or macronutrient intake values were in the top or bottom 1% of all participants energy and macronutrient intake data from that questionnaire
- their dietary data from that questionnaire was coded by UK Biobank as "not credible" based on their energy intake
- their energy intake was less than 1.1 × their basal metabolic rate[35]
- they said that their diet yesterday wasn't typical for them

A participant's data from all diet questionnaires was excluded if, when asked at the baseline clinic visit or the first repeat visit:

- they said that they'd made a major change to their diet in the last 5 years
- they said that their diet varies much from week to week (or said that they did not know or preferred not to answer)
- they said that they were current smokers (or said that they preferred not to answer)

Participants were also excluded unless their BMI was between 19.5 and 34.5kg/m^2. These cut-offs were chosen since there were less than 500 participants in each 0.5kg/m^2 interval outside of this range, and this could increase uncertainty in any diet-BMI models fitted in later analyses.

The remaining energy and macronutrient data from the baseline visit questionnaire and each of the online questionnaires was averaged to create estimated average dietary intake variables for each participant.

4.2.1.1 Observational analyses in UK Biobank

The relationship between dietary intake and BMI in UK Biobank was explored. 14 energy and macronutrient variables were studied: total energy intake (kJ/d), protein intake (g/d and % total energy intake), fat intake (g/d and % total energy intake), carbohydrate intake (g/d and % total energy intake), saturated fat intake (g/d and % total energy intake), polyunsaturated fat intake (g/d and % total energy intake), total sugar intake (g/d and % total energy intake) and fibre intake (g/d).

Linear regression models were fitted to explore the relationship between macronutrient intake and BMI in UK Biobank. Models were adjusted for age (at the baseline assessment visit) and sex. Analyses were performed in R (version 3.3.3) using the *lm* function from the *stats* package. Models were fitted for diet → BMI since this is the more intuitive relationship between diet and BMI. Models were also fitted for BMI → diet to enable comparison later with results from BMI allele score → diet analyses.

```
lm(BMI ~ diet + age + sex)
```

```
lm(diet ~ BMI + age + sex)
```

Separate models for females and males were also fitted.

4.2.1.2 Macronutrient intake to BMI analyses – two-sample MR

As discussed in the previous chapter, some genetic instruments for macronutrient intake have been identified in UK Biobank. Heritability estimates from the UK Biobank macronutrient data are low (c.3-6%, **Table 8**). Therefore, a large sample size is needed when using these macronutrient SNPs as instruments for diet in MR analyses.

In situations, such as this, where it is labour-intensive and expensive to obtain a sample in which both the exposure and outcome variables are available, two-sample MR can be used to investigate causality between the exposure and outcome (**2.2.4.1**).[16,128] In two-sample MR, the instrument-exposure and instrument-outcome coefficients are obtained

from summary data from a GWAS of the exposure variable and a GWAS (in a different sample) of the outcome variable respectively.

Two-sample MR analyses were performed to estimate the causal effect of various macronutrients on BMI. The instrument-exposure coefficients are taken from the UK Biobank macronutrient GWAS in the previous chapter (**Table 4**; results from models without adjustment for BMI), and the instrument-outcome coefficients are from a BMI GWAS conducted in c.320,000 individuals of European descent (**2.2.4**).[52]

These two-sample MR analyses were conducted in R (version 3.3.3) using the *mr_singlesnp* function from the *TwoSampleMR* package.[131] SNPs that reached genome-wide significance in the "online group" of the diet GWAS were used as genetic instruments, regardless of whether they replicated in the "visit" group (**3.3.2**). If results for a SNP were not available in the BMI GWAS summary results, then a proxy (r^2>0.6) was used. No suitable proxies were available for rs13447258 or rs200553669. Most of the macronutrients only had a single SNP available to use as an instrument, so the Wald ratio was used to calculate the MR estimates.[55] If there was more than one instrumental SNP for a macronutrient then IVW regression was also performed.[129,130]

4.2.1.3 BMI to macronutrient intake analyses in UK Biobank

The next objective was to explore the causal effect of BMI on dietary behaviour. The 97-SNP BMI score from Locke et al., which is often used as a genetic instrument for BMI, is not suitable for MR analyses of BMI on diet, since several of the 97 SNPs are thought to influence BMI through dietary choices.[52] For example, the association observed between *FTO* SNPs and BMI appears to be, in part, due to the association between *FTO* SNPs and appetite.[208] This violates the requirement that, in MR analysis, the genetic instrument should only be associated with the outcome variable through its relationship with the exposure variable (**2.2.4**).

Locke et al. reviewed literature on 405 genes that are within 500 kB and r^2>0.2 of their 97 SNPs, and used this information to classify the genes into one or more biological categories (Locke et al. Supplementary Table 22).[52] 25 of the categories contained at

least three genes. Each gene (and hence each SNP) could appear in more than one of these biological categories, so there is some overlap between the categories (**Table 10**). A leave-one-out (LOO) weighted allele score was created for each category to explore whether the association between the 97 SNP BMI score and diet in UK Biobank is driven purely by a single aspect of the score (e.g. SNPs in the hypothalamic expression and regulatory function category). For each category, the LOO weighted allele score was created using all the SNPs (from the 97 SNPs) except for SNPs in that category. The weights for these allele scores were taken from the European sex-combined analysis (Locke et al. Supplementary Table 4).

The association between each of the LOO allele scores and the UK Biobank macronutrient intake variables was tested by fitting the model:

```
lm(macronutrient intake ~ LOO allele score + age + sex)
```

Separate models for females and males were also fitted. Analyses were performed in R (version 3.3.3) using the *lm* function from the *stats* package (**2.2.1**).

4.2.2. Dietary patterns and BMI analyses

4.2.2.1 Observational analyses in ALSPAC

Dietary behaviour was measured using PCs previously generated from FFQs and diet diaries (**2.1.1.1**).[27,101,209]

Multivariable linear regression was performed for each of the diet PCs separately on BMI in the ALSPAC children at ages 7 and 13 years. Three different models were fitted: a model adjusted for age at BMI measurement and sex; a model adjusted for age, sex and maternal education; and (for the diet diary PCs only) a model adjusted for total energy intake, age and sex. Maternal education was summarised as a binary variable indicating whether the mother had completed A Levels (and/or university) or not. Daily energy intake (estimated from the diet diaries) was included as a covariate in the third model as

studies have suggested that body weight is most affected by total energy consumption rather than proportions of different macronutrients consumed.[202]

```
lm(BMI ~ diet PC + age + sex)

lm(BMI ~ diet PC + age + sex + maternal education)

lm(BMI ~ diet PC + age + sex + energy intake)
```

Models were also fitted for BMI → diet to enable comparison later with results from BMI allele score → diet analyses.

```
lm(diet PC ~ BMI + age + sex)

lm(diet PC ~ BMI + age + sex + maternal education)

lm(diet PC ~ BMI + age + sex + energy intake)
```

Participants were excluded from the analyses at age 7 if their BMI at that age was not between 13 and 21.5kg/m^2. Participants were excluded from the analyses at age 15 if their BMI at that age was not between 15 and 28kg/m^2. These cut-offs were chosen since there were less than 50 participants in each 0.5kg/m^2 interval outside of these ranges.

4.2.2.2 Diet PCs to BMI

Investigating the causal effect of diet on BMI is challenging as a suitable genetic instrument for diet is needed to perform MR. The diet PCs used in the cross-sectional analyses above are unique to the FFQ and diet diary data used to generate those PCs, and no genetic instruments have been identified for them. To avoid overfitting, genetic variants should be obtained from a different sample to the sample in which MR is performed.[210,211] Sometimes samples are split in two to facilitate this,[212] however this would not be practical here since a large sample size is needed to identify genetic variants for complex traits such as dietary behaviour.

4.2.2.3 BMI allele scores to diet PCs analyses in ALSPAC

The association between each of the LOO allele scores and the ALSPAC diet PCs was examined by fitting the model:

```
lm(diet PC ~ LOO allele score + age + sex)
```

Diet PCs were only studied if there was suggestive evidence of an association with BMI in the cross-sectional analyses.

Table 10 – BMI SNPs grouped by functional category.

SNPs are in bold if they explained at least 0.05% variance in BMI in the European sex-combined analysis in the Locke et al. GWAS.

Biological subcategory	SNPs
Neuronal Developmental processes	**rs3101336**, rs7138803, rs3888190, rs2287019, rs16951275, rs3810291, rs7141420, rs13078960, rs12286929, rs11165643, rs4256980, rs17094222, rs2820292, rs12885454, rs6804842, rs4740619, rs492400, rs13191362, rs3736485, rs2080454, rs2033529, rs7239883, rs2836754, rs9400239, rs10733682, rs11057405, rs29941, rs4787491, rs13201877
Neurotransmission	**rs10938397**, rs10132280, rs1167827, rs9925964, rs13191362, rs3736485, rs9914578, rs4787491, rs7899106, rs9540493
Hypothalamic expression and regulatory function	**rs1558902**, **rs6567160**, **rs13021737**, **rs10938397**, **rs11030104**, rs10182181, rs3888190, rs1516725, rs17405819, rs4256980, rs7164727, rs3736485, rs7243357
Neuronal Expression	rs10182181, rs12446632, rs10968576, rs12401738, rs7599312, rs11126666, rs13191362, rs2075650, rs11583200, rs9914578, rs3849570, rs1808579
Lipid biosynthesis and metabolism	rs3817334, rs2112347, rs1928295, rs2650492, rs7164727, rs492400, rs11191560, rs2075650, rs4787491, rs1808579
Bone Development	rs16951275, rs6091540, rs205262, rs9641123, rs16851483, rs1167827, rs6804842, rs11727676, rs17724992
Signalling (includes subcategories)	**rs6567160**, **rs11030104**, rs10182181, rs3888190, rs12446632, rs2287019, rs16951275, rs3817334, rs12566985, rs17024393, rs7903146, rs4256980, rs17094222, rs12401738, rs7599312, rs9641123, rs16851483, rs12940622, rs7239883, rs4787491, rs17203016, rs7243357
GPCR	rs6091540, rs205262, rs492400, rs3736485, rs17724992
Mitochondrial	rs3888190, rs3817334, rs3810291, rs13107325, rs17094222, rs12016871, rs9925964, rs2075650
Retinoic Acid Receptors	rs12446632, rs3817334, rs12429545, rs17094222, rs6804842, rs492400
Endocytosis/Exocytosis	**rs543874**, rs10132280, rs17094222, rs1167827, rs9925964, rs13191362, rs3736485, rs17001654, rs1000940, rs7239883, rs11688816, rs11057405, rs2121279, rs1808579
Eye-related	rs17024393, rs17094222, rs2820292, rs7164727, rs10733682
Tumorigenesis	rs7138803, rs3817334, rs12429545, rs4256980, rs7599312, rs16851483, rs3736485, rs2836754, rs11688816, rs2121279, rs4787491
Apoptosis	rs3817334, rs17094222, rs2365389, rs1167827, rs758747, rs9925964, rs2650492, rs4740619, rs13191362, rs1000940, rs2033529, rs2836754, rs11057405
Membrane Proteins	**rs6567160**, **rs13021737**, rs3817334, rs2112347, rs3810291, rs7599312, rs1167827, rs2075650, rs1000940, rs2033529, rs29941
Hormone metabolism/regulation	rs10182181, rs3888190, rs2176598, rs17203016
Purine/Pyrimidine	rs17024393, rs2365389, rs11191560, rs977747
Monogenic Obesity and/or Energy Homeostasis	**rs1558902**, **rs6567160**, **rs11030104**, rs10182181, rs3888190, rs4256980, rs7164727, rs3736485, rs6465468
Immune system	rs3817334, rs2112347, rs13078960, rs12286929, rs17094222, rs12401738, rs12885454, rs9641123, rs758747, rs1928295, rs9925964, rs11847697, rs2075650, rs9374842, rs13201877
Limb development	**rs2207139**, rs6804842, rs10733682
Ubiquitin pathways	rs1016287, rs12401738, rs205262, rs9925964, rs13191362, rs1528435
Glucose homeostasis and/or diabetes	**rs6567160**, rs2287019, rs3817334, rs7903146, rs12940622, rs7164727, rs2176040, rs11583200, rs9400239, rs3849570, rs17203016
Cell cycle	**rs1558902**, rs10182181, rs2112347, rs657452, rs1016287, rs4256980, rs12401738, rs12885454, rs1167827, rs758747, rs9925964, rs7164727, rs11847697, rs492400, rs1000940, rs11057405, rs9914578, rs977747, rs9374842, rs4787491, rs2245368, rs1808579, rs1460676
DNA repair (nuclear trafficking)	rs4256980, rs2820292, rs1167827, rs17001654
Muscle biology	**rs10938397**, rs3888190, rs3817334, rs9925964, rs3849570, rs4787491

4.3. Results

4.3.1. Macronutrient intake and BMI results

4.3.1.1 Results from observational analyses in UK Biobank

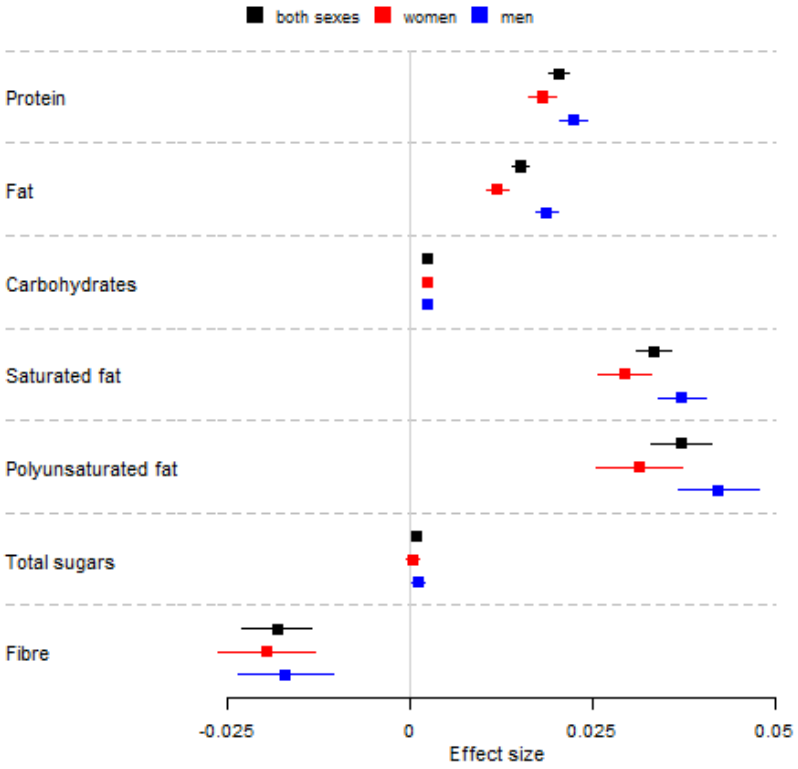
Strong positive associations were observed between BMI and g/d intake of protein, fat, carbohydrates, saturated fat and polyunsaturated fat (**Figure 7a**; **Table 11**; **Table 12**). The effect sizes were larger in men than women. A positive association was also observed between BMI and g/d intake of sugar in men but not women. Fibre intake (g/d) was negatively associated with BMI.

In analyses where macronutrients were quantified as percentage of total energy intake, BMI was positively associated with fat, saturated fat and polyunsaturated fat intake, and effect estimates were similar in men and women (**Figure 7b**; **Table 11**; **Table 12;**). BMI was negatively associated with carbohydrate and total sugar intake, and these effect estimates were much larger in men than women. BMI was positively associated with protein intake in women but not men.

Figure 7 – Forest plots of diet → BMI associations.

Using average (from all five questionnaires) diet variables. BMI is measured in kg/m^2.

(a) Macronutrients quantified as g/day



(b) Macronutrients quantified as % energy intake.

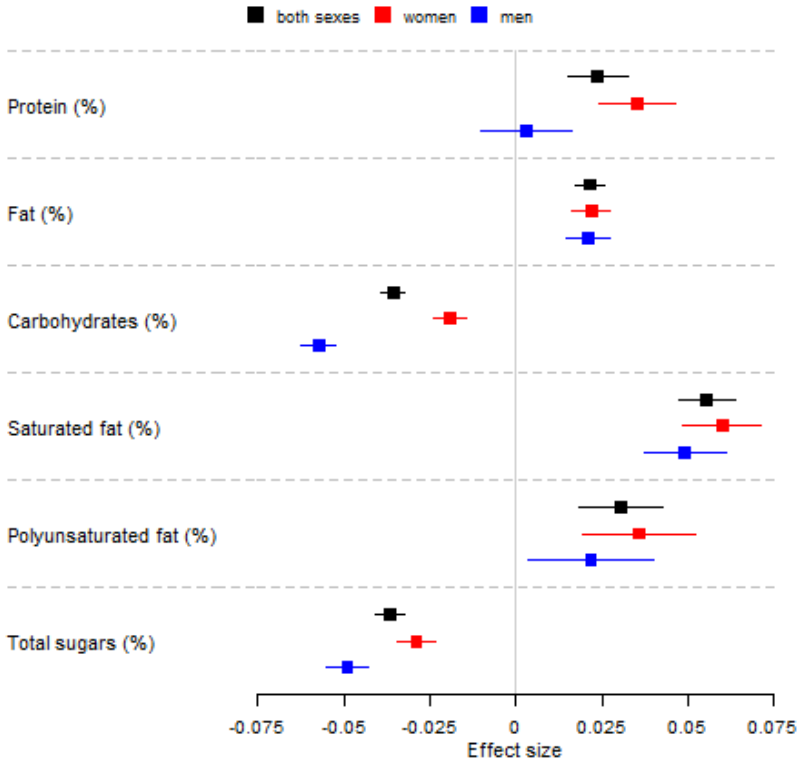


Table 11 – Diet → BMI associations.

Using average (from all five questionnaires) diet variables. BMI is measured in kg/m^2. N=54,100 in analyses with both sexes, N=31,169 in female only analyses and N=22,931 in male only analyses.

	Both sexes			Female only			Male only		
	Beta	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
Energy (kJ/d)	2.60×10^{-4}	2.44×10^{-4}, 2.75×10^{-4}	1.43×10^{-234}	1.77×10^{-4}	1.56×10^{-4}, 1.99×10^{-4}	3.38×10^{-57}	3.56×10^{-4}	3.35×10^{-4}, 3.78×10^{-4}	4.30×10^{-223}
Protein (g/d)	0.020	0.019, 0.022	2.77×10^{-194}	0.018	0.016, 0.02	3.63×10^{-76}	0.022	0.02, 0.024	9.13×10^{-126}
Fat (g/d)	0.015	0.014, 0.016	1.00×10^{-157}	0.012	0.01, 0.013	1.32×10^{-49}	0.019	0.017, 0.02	3.54×10^{-124}
Carbohydrates (g/d)	0.002	0.002, 0.003	1.46×10^{-28}	0.002	0.002, 0.003	6.09×10^{-15}	0.002	0.002, 0.003	3.23×10^{-14}
Saturated fat (g/d)	0.033	0.031, 0.036	2.46×10^{-160}	0.029	0.026, 0.033	5.33×10^{-60}	0.037	0.034, 0.04	3.69×10^{-109}
Polyunsat. fat (g/d)	0.037	0.033, 0.041	1.28×10^{-69}	0.031	0.025, 0.037	2.10×10^{-25}	0.042	0.037, 0.048	4.36×10^{-49}
Total sugars (g/d)	0.001	0.0002, 0.001	0.01	0.0004	-0.001, 0.001	0.38	0.001	0.0003, 0.002	0.01
Fibre (g/d)	-0.018	-0.023, -0.014	1.29×10^{-14}	-0.020	-0.026, -0.013	3.86×10^{-9}	-0.017	-0.024, -0.011	2.36×10^{-7}
Protein (%)	0.024	0.015, 0.032	5.02×10^{-8}	0.035	0.024, 0.047	6.32×10^{-10}	0.003	-0.01, 0.016	0.642248
Fat (%)	0.022	0.017, 0.026	2.36×10^{-23}	0.022	0.016, 0.028	6.28×10^{-14}	0.021	0.014, 0.027	1.10×10^{-10}
Carbohydrates (%)	-0.036	-0.039, -0.032	4.25×10^{-90}	-0.019	-0.024, -0.015	1.80×10^{-15}	-0.057	-0.062, -0.052	1.83×10^{-112}
Saturated fat (%)	0.056	0.047, 0.064	4.04×10^{-40}	0.060	0.049, 0.071	1.22×10^{-25}	0.049	0.037, 0.061	6.24×10^{-16}
Polyunsat. fat (%)	0.030	0.018, 0.042	9.46×10^{-7}	0.036	0.019, 0.052	1.72×10^{-5}	0.022	0.004, 0.04	0.02
Total sugars (%)	-0.037	-0.041, -0.033	2.59×10^{-69}	-0.029	-0.034, -0.023	2.90×10^{-25}	-0.049	-0.055, -0.043	1.31×10^{-55}

Table 12 – BMI → diet associations.

Using average (from all five questionnaires) diet variables. BMI is measured in kg/m^2. N=54,100 in analyses with both sexes, N=31,169 in female only analyses and N=22,931 in male only analyses.

	Both sexes			Female only			Male only		
	Beta	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
Energy (kJ/d)	75.3	70.8, 79.8	1.4×10^{-234}	45.8	40.1, 51.4	3.4×10^{-57}	121.7	114.3, 129.1	4.3×10^{-223}
Protein (g/d)	0.80	0.75, 0.85	2.8×10^{-194}	0.60	0.54, 0.67	3.6×10^{-76}	1.10	1.01, 1.19	9.1×10^{-126}
Fat (g/d)	0.87	0.80, 0.93	1.0×10^{-157}	0.59	0.51, 0.67	1.3×10^{-49}	1.30	1.19, 1.41	3.5×10^{-124}
Carbohydrates (g/d)	0.96	0.79, 1.12	1.5×10^{-28}	0.82	0.62, 1.03	6.1×10^{-15}	1.11	0.83, 1.40	3.2×10^{-14}
Saturated fat (g/d)	0.40	0.37, 0.43	2.5×10^{-160}	0.29	0.26, 0.33	5.3×10^{-60}	0.57	0.52, 0.62	3.7×10^{-109}
Polyunsat. fat (g/d)	0.16	0.14, 0.17	1.3×10^{-69}	0.11	0.09, 0.13	2.1×10^{-25}	0.22	0.19, 0.25	4.4×10^{-49}
Total sugars (g/d)	0.14	0.03, 0.26	0.01	0.06	-0.07, 0.20	0.38	0.25	0.06, 0.44	0.01
Fibre (g/d)	-0.06	-0.08, -0.04	1.3×10^{-14}	-0.06	-0.08, -0.04	3.9×10^{-9}	-0.07	-0.09, -0.04	2.4×10^{-7}
Protein (%)	0.02	0.01, 0.03	5.0×10^{-8}	0.03	0.02, 0.05	6.3×10^{-10}	0.00	-0.01, 0.02	0.64
Fat (%)	0.08	0.07, 0.10	2.4×10^{-23}	0.08	0.06, 0.10	6.3×10^{-14}	0.09	0.06, 0.11	1.1×10^{-10}
Carbohydrates (%)	-0.21	-0.23, -0.19	4.2×10^{-90}	-0.11	-0.13, -0.08	1.8×10^{-15}	-0.38	-0.42, -0.35	1.8×10^{-112}
Saturated fat (%)	0.06	0.05, 0.07	4.0×10^{-40}	0.06	0.05, 0.07	1.2×10^{-25}	0.06	0.04, 0.07	6.2×10^{-16}
Polyunsat. fat (%)	0.01	0.01, 0.02	9.5×10^{-7}	0.02	0.01, 0.02	1.7×10^{-5}	0.01	0.002, 0.02	0.02
Total sugars (%)	-0.16	-0.17, -0.14	2.6×10^{-69}	-0.12	-0.14, -0.10	2.9×10^{-25}	-0.22	-0.25, -0.19	1.3×10^{-55}

4.3.1.2 Macronutrient intake to BMI - two-sample MR results

Two-sample MR analyses were conducted to investigate the causal effect of dietary intake on BMI. Results of these analyses are given in **Figure 8** and **Table 13**. These results suggest that a 1g/d increase in protein intake leads to a decrease in BMI of 0.014kg/m^2 (95% CI 0.002, 0.026), and a 1g/d increase in polyunsaturated fat intake leads to a decrease in BMI of 0.043kg/m^2 (95% CI 0.004, 0.082). There is also weaker evidence suggesting that a 1g/d increase in fat intake leads to a 0.009kg/m^2 (-0.003, 0.021) decrease in BMI, and a 1g/d increase in saturated fat intake leads to a 0.021kg/m^2 (-0.008, 0.050) decrease in BMI.

Figure 8 – Forest plot of results from diet → BMI two-sample MR analyses.
Effect sizes are the increase in BMI (kg/m^2) per 1g/d increase in macronutrient intake.

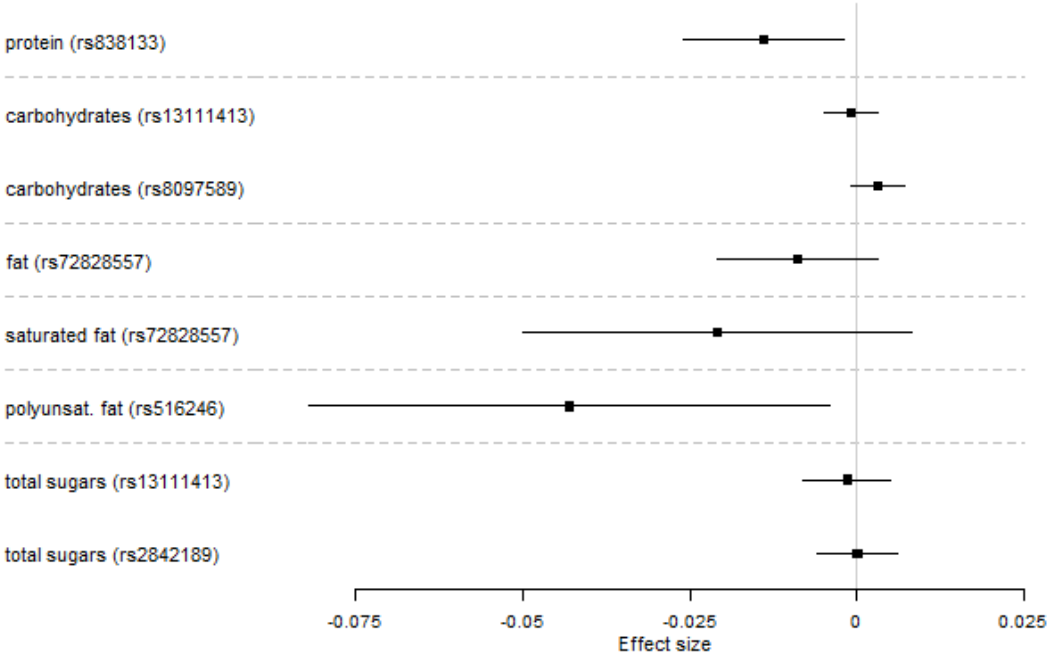


Table 13 – Results from two-sample MR analyses investigating the causal effect of diet on BMI.

MR estimates are calculated using the Wald ratio for single SNPs, and IVW where there is more than one SNP. Energy is measured in kJ/d, macronutrients are measured in g/d, BMI is measured in kg/m^2. LD estimates from 1000 Genomes European populations. *SNPs that replicated in diet GWAS in previous chapter.

Trait	SNP	EA	SNP available in GIANT GWAS?	Beta	95% CI	p-value
Energy	rs7957145	T	Yes.	-6.53×10^{-6}	-1.45×10^{-4}, 1.32×10^{-4}	0.926
Protein	rs838133*	G	Yes.	-0.014	-0.026, -0.002	0.022
	rs13447258	A	No. No suitable proxy.	-	-	-
Carbs	rs13111413	T	Yes.	-0.001	-0.005, 0.003	0.640
	rs8097589	A	Yes.	0.003	-0.001, 0.007	0.209
			IVW MR	0.001	-0.003, 0.004	0.670
Fat	rs72828557*	T	No. Used rs17568354 as a proxy (R^2=0.663).	-0.009	-0.021, 0.003	0.158
Saturated fat	rs72828557*	T	No. Used rs17568354 as a proxy (R^2=0.663).	-0.021	-0.050, 0.008	0.158
Polyunsaturated fat	rs516246*	T	Yes.	-0.043	-0.082, -0.004	0.029
Fibre	rs200553669	G	No. No suitable proxy.	-	-	-
Total sugars	rs13111413	T	Yes.	-0.001	-0.008, 0.005	0.640
	rs2842189*	C	No. Used rs2782640 as a proxy (LD=0.974).	-8.57×10^{-5}	-0.006, 0.006	0.979
			IVW MR	-0.001	-0.005, 0.004	0.724

4.3.1.3 BMI allele score and macronutrient intake results from UK Biobank

The relationship between macronutrient intake and the GIANT BMI score and each of the LOO weighted allele scores was explored (both sexes, women only, men only) (**Figure 9 a, b, c** respectively). The strongest associations observed were positive associations between the allele scores and fibre intake (g/d), mainly driven by the men. The allele scores were positively associated with protein intake (g/d and % energy intake) in the women. Weak suggestive evidence of positive associations between the allele scores and intake of energy (kJ/d), fat (g/d), carbohydrates (g/d) and polyunsaturated fat (g/d) was observed in the men.

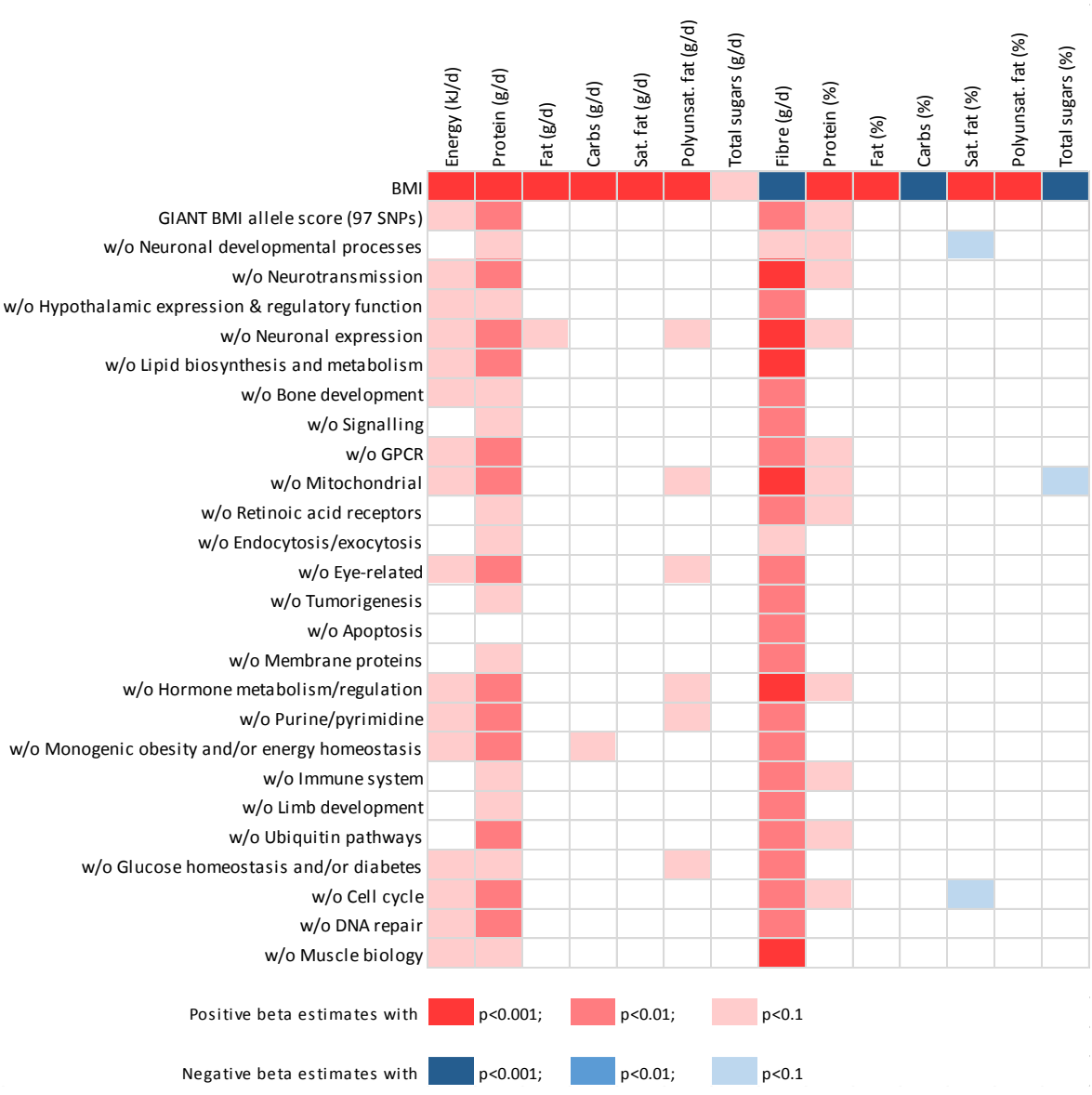
The observed effect estimates were mostly directionally consistent with those between BMI and the macronutrients, except for fibre intake (g/d) which was negatively associated with BMI, but positively associated with the allele scores.

Generally, the strengths of associations between a macronutrient and the allele scores were fairly consistent. However, associations tended to be weaker for the allele scores without SNPs in the neuronal developmental processes, hypothalamic expression and regulatory function, and endocytosis/exocytosis categories.

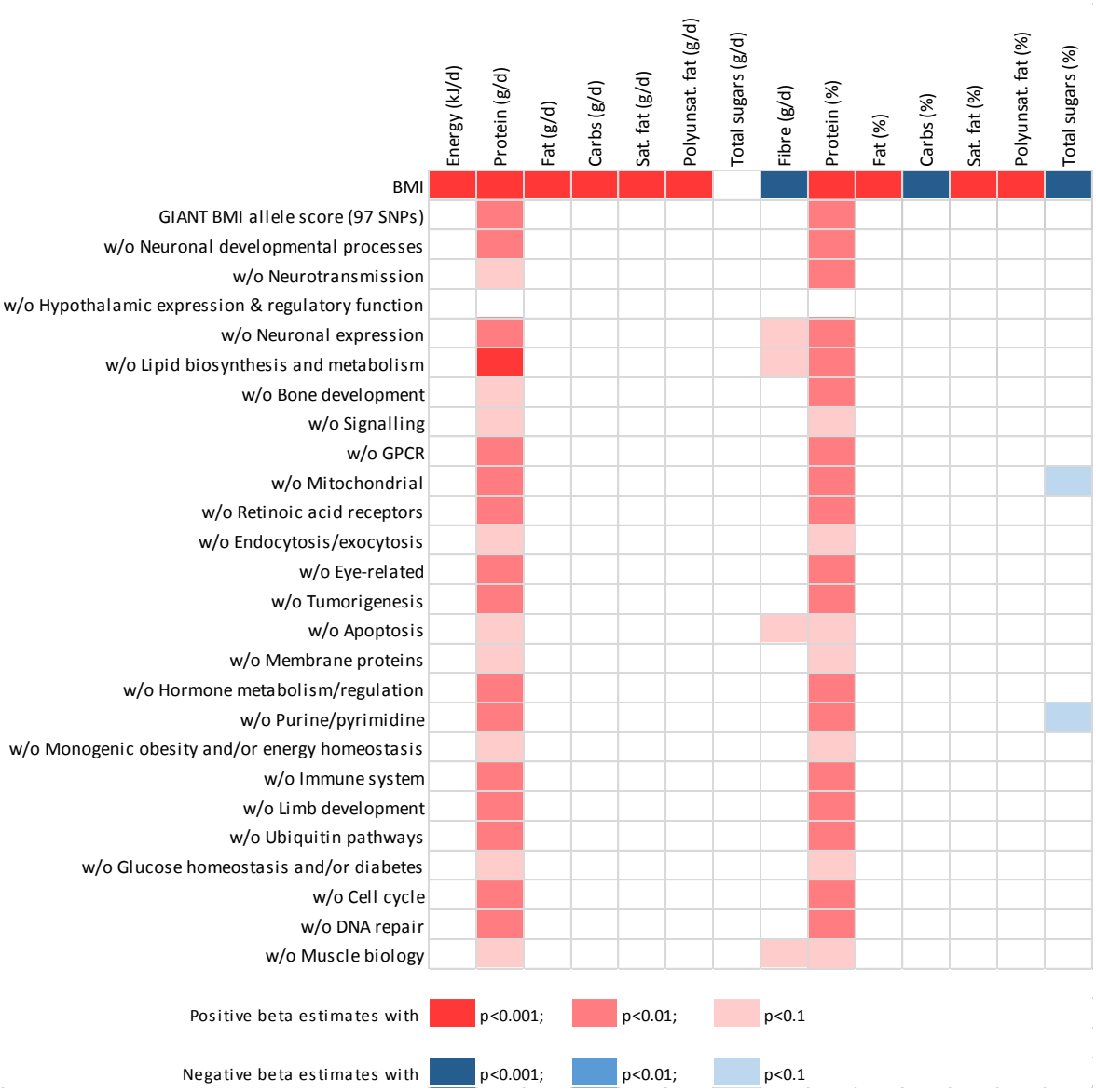
Figure 9 – Heatmap showing the effect strengths and directions from the relationships between the BMI allele scores and macronutrient intake.

p-values and effect directions are from the model: macronutrient intake ~ BMI allele score + age + sex; N=27,459 for both sexes, N=15,847 for women, N=11,612 for men. The relationship between BMI and macronutrient intake is also provided for comparison with the allele scores; p-values and effect directions are from the model: macronutrient intake ~ BMI + age + sex. Only people with BMIs between 19.5kg/m^2 and 34.5kg/m^2 were included in the analyses.

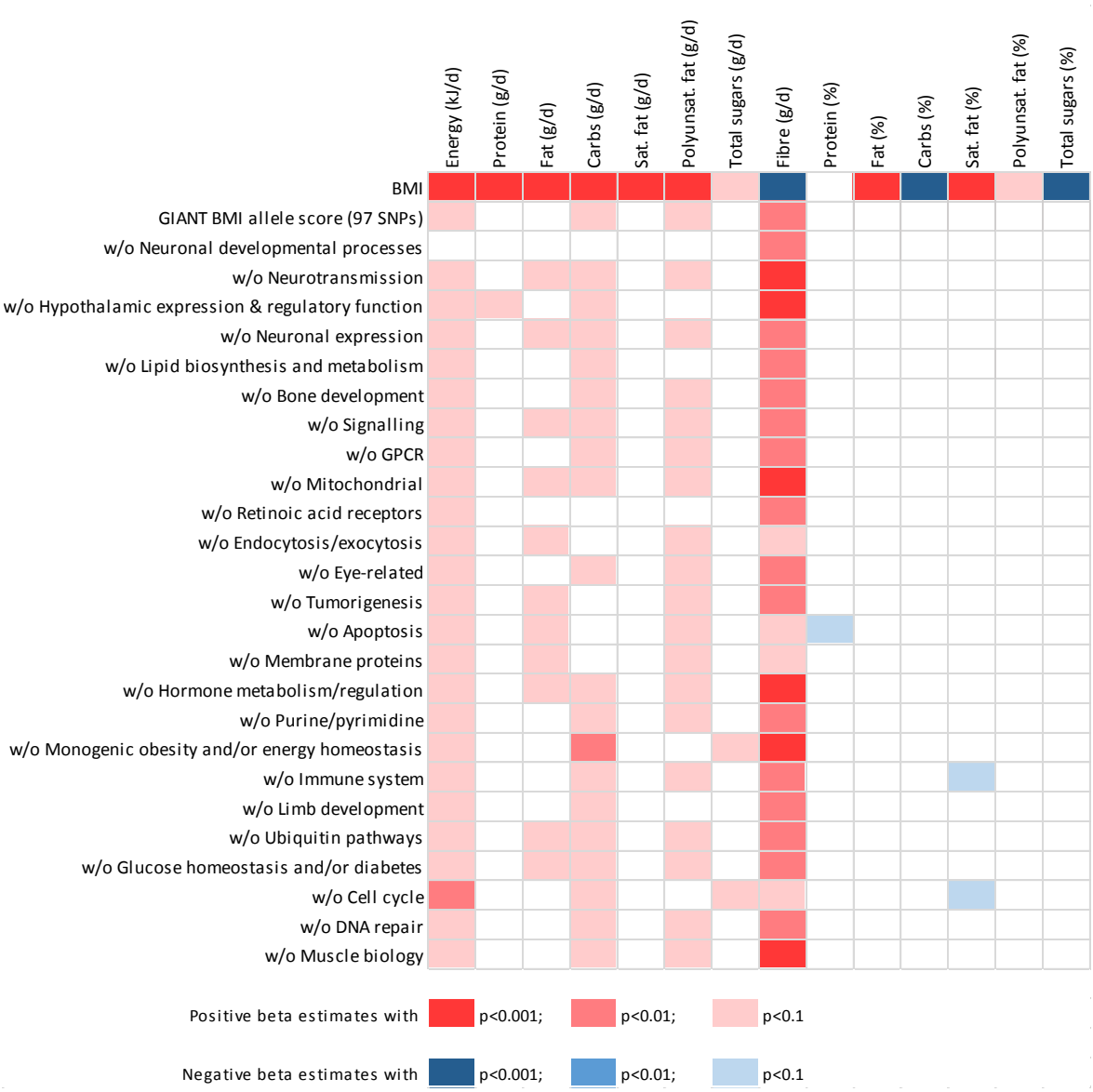
(a) Both sexes



(b) Women only


Energy (kJ/d)
Protein (g/d)
Fat (g/d)
Carbs (g/d)
Sat. fat (g/d)
Polyunsat. fat (g/d)
Total sugars (g/d)
Fibre (g/d)
Protein (%)
Fat (%)
Carbs (%)
Sat. fat (%)
Polyunsat. fat (%)
Total sugars (%)
BMI
GIANT BMI allele score (97 SNPs)
w/o Neuronal developmental processes
w/o Neurotransmission
w/o Hypothalamic expression & regulatory function
w/o Neuronal expression
w/o Lipid biosynthesis and metabolism
w/o Bone development
w/o Signalling
w/o GPCR
w/o Mitochondrial
w/o Retinoic acid receptors
w/o Endocytosis/exocytosis
w/o Eye-related
w/o Tumorigenesis
w/o Apoptosis
w/o Membrane proteins
w/o Hormone metabolism/regulation
w/o Purine/pyrimidine
w/o Monogenic obesity and/or energy homeostasis
w/o Immune system
w/o Limb development
w/o Ubiquitin pathways
w/o Glucose homeostasis and/or diabetes
w/o Cell cycle
w/o DNA repair
w/o Muscle biology
Positive beta estimates with p<0.001; p<0.01; p<0.1
Negative beta estimates with p<0.001; p<0.01; p<0.1

(c) Men only


Energy (kJ/d)
Protein (g/d)
Fat (g/d)
Carbs (g/d)
Sat. fat (g/d)
Polyunsat. fat (g/d)
Total sugars (g/d)
Fibre (g/d)
Protein (%)
Fat (%)
Carbs (%)
Sat. fat (%)
Polyunsat. fat (%)
Total sugars (%)
BMI
GIANT BMI allele score (97 SNPs)
w/o Neuronal developmental processes
w/o Neurotransmission
w/o Hypothalamic expression & regulatory function
w/o Neuronal expression
w/o Lipid biosynthesis and metabolism
w/o Bone development
w/o Signalling
w/o GPCR
w/o Mitochondrial
w/o Retinoic acid receptors
w/o Endocytosis/exocytosis
w/o Eye-related
w/o Tumorigenesis
w/o Apoptosis
w/o Membrane proteins
w/o Hormone metabolism/regulation
w/o Purine/pyrimidine
w/o Monogenic obesity and/or energy homeostasis
w/o Immune system
w/o Limb development
w/o Ubiquitin pathways
w/o Glucose homeostasis and/or diabetes
w/o Cell cycle
w/o DNA repair
w/o Muscle biology
Positive beta estimates with p<0.001; p<0.01; p<0.1
Negative beta estimates with p<0.001; p<0.01; p<0.1

4.3.2. Diet PCs and BMI results

4.3.2.1 Results from dietary patterns and BMI analyses in ALSPAC

Cross-sectional analyses were performed to explore the relationship between the diet PCs and BMI at ages 7 and 13 years. Three models were fitted: model 1 was adjusted for age and sex; model 2 was adjusted for age, sex and maternal education; and model 3 was adjusted for age, sex and energy intake (**Figure 10**; **Table 14**; **Table 15**).

Several of the PCs showed evidence of an association with BMI. The strongest associations were a positive association with the age 7 diet diary "packed lunch" PC, and negative associations with the age 13 FFQ "snack/sugared drinks" PC, the age 13 diet diary "traditional" PC, and the age 7 diet diary "health aware" PC.

The age 13 FFQ "traditional/health conscious" PC showed a suggestive negative association with BMI, which attenuated after adjustment for maternal education. The age 13 diet diary "packed lunch" PC was positively associated with BMI in the model adjusted for energy intake.

Figure 10 - Forest plot of diet → BMI observational analyses.

Model 1 is adjusted for age and sex; model 2 is adjusted for age, sex and maternal education; model 3 is adjusted for age, sex and energy intake. Effect sizes are the increase in BMI per unit increase in diet PC score. DD, diet diary.

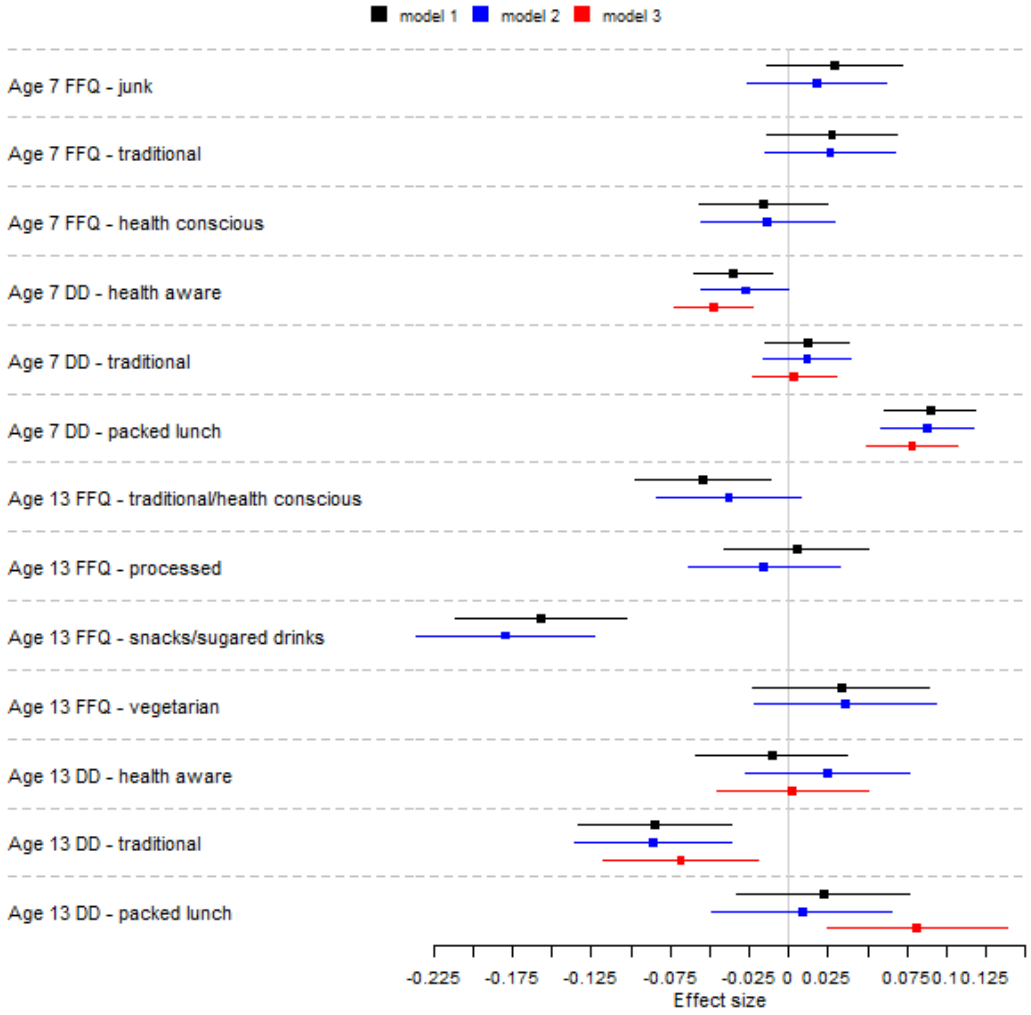


Table 14 – Results from diet → BMI cross-sectional analyses.

DD, diet diary.

	Adjusted for age and sex			Adjusted for age, sex and mat. ed.			Adjusted for age, sex and energy intake		
	Beta	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
Age 7 – FFQs		(N=6,171)			(N=6,058)				
FFQ PC 1 "junk"	0.029	-0.013, 0.072	0.179	0.018	-0.026, 0.062	0.430	-	-	-
FFQ PC 2 "traditional"	0.027	-0.014, 0.068	0.194	0.026	-0.015, 0.068	0.215	-	-	-
FFQ PC 3 "health conscious"	-0.016	-0.057, 0.025	0.437	-0.013	-0.055, 0.029	0.535	-	-	-
Age 7 – diet diaries		(N=6,862)			(N=6,308)			(N=6,862)	
DD PC 1 "health aware"	-0.035	-0.061, -0.01	0.006	-0.028	-0.055, -0.0001	0.049	-0.048	-0.073, -0.022	2.11×10^{-4}
DD PC 2 "traditional"	0.012	-0.015, 0.039	0.382	0.011	-0.016, 0.039	0.413	0.003	-0.023, 0.03	0.799
DD PC 3 "packed lunch"	0.089	0.061, 0.118	6.26×10^{-10}	0.088	0.058, 0.117	6.15×10^{-9}	0.078	0.05, 0.106	6.27×10^{-8}
Age 13 – FFQs		(N=3,898)			(N=3,661)				
FFQ PC 1 "traditional/health conscious"	-0.055	-0.098, -0.012	0.013	-0.038	-0.084, 0.008	0.102	-	-	-
FFQ PC 2 "processed"	0.005	-0.041, 0.05	0.832	-0.016	-0.064, 0.032	0.512	-	-	-
FFQ PC 3 "snacks/sugared drinks"	-0.157	-0.211, -0.103	1.20×10^{-8}	-0.180	-0.236, -0.123	5.49×10^{-10}	-	-	-
FFQ PC 4 "vegetarian"	0.033	-0.022, 0.089	0.241	0.036	-0.022, 0.094	0.220	-	-	-
Age 13 – diet diaries		(N=5,691)			(N=5,269)			(N=5,691)	
DD PC 1 "health aware"	-0.011	-0.059, 0.037	0.664	0.025	-0.027, 0.076	0.350	0.002	-0.046, 0.05	0.927
DD PC 2 "traditional"	-0.085	-0.133, -0.036	0.001	-0.086	-0.136, -0.036	0.001	-0.069	-0.117, -0.02	0.006
DD PC 3 "packed lunch"	0.022	-0.033, 0.076	0.434	0.008	-0.049, 0.066	0.772	0.081	0.024, 0.138	0.005

Table 15 - Results from BMI → diet cross-sectional analyses.

DD, diet diary.

	Adjusted for age and sex			Adjusted for age, sex and mat. ed.			Adjusted for age, sex and energy intake		
	Beta	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
Age 7 – FFQs		(N=6,171)			(N=6,058)				
FFQ PC 1 "junk"	0.010	-0.005, 0.025	0.179	0.006	-0.009, 0.02	0.430	-	-	-
FFQ PC 2 "traditional"	0.010	-0.005, 0.025	0.194	0.010	-0.006, 0.025	0.215	-	-	-
FFQ PC 3 "health conscious"	-0.006	-0.021, 0.009	0.437	-0.005	-0.02, 0.01	0.535	-	-	-
Age 7 – diet diaries		(N=6,862)			(N=6,308)			(N=6,862)	
DD PC 1 "health aware"	-0.031	-0.053, -0.009	0.006	-0.022	-0.044, 0	0.049	-0.042	-0.064, -0.02	2.11×10^{-4}
DD PC 2 "traditional"	0.009	-0.012, 0.03	0.382	0.009	-0.013, 0.031	0.413	0.003	-0.018, 0.024	0.799
DD PC 3 "packed lunch"	0.062	0.043, 0.082	6.26×10^{-10}	0.061	0.04, 0.081	6.15×10^{-9}	0.055	0.035, 0.074	6.27×10^{-8}
Age 13 – FFQs		(N=3,898)			(N=3,661)				
FFQ PC 1 "traditional/health conscious"	-0.029	-0.052, -0.006	0.013	-0.019	-0.042, 0.004	0.102	-	-	-
FFQ PC 2 "processed"	0.002	-0.019, 0.024	0.832	-0.007	-0.029, 0.015	0.512	-	-	-
FFQ PC 3 "snacks/sugared drinks"	-0.053	-0.071, -0.035	1.20×10^{-8}	-0.058	-0.077, -0.04	5.49×10^{-10}	-	-	-
FFQ PC 4 "vegetarian"	0.011	-0.007, 0.028	0.241	0.011	-0.007, 0.03	0.220	-	-	-
Age 13 – diet diaries		(N=5,691)			(N=5,269)			(N=5,691)	
DD PC 1 "health aware"	-0.003	-0.017, 0.011	0.664	0.007	-0.007, 0.021	0.350	0.001	-0.013, 0.015	0.927
DD PC 2 "traditional"	-0.024	-0.038, -0.01	6.38×10^{-4}	-0.025	-0.04, -0.01	7.71×10^{-4}	-0.019	-0.033, -0.006	0.006
DD PC 3 "packed lunch"	0.005	-0.007, 0.017	0.434	0.002	-0.011, 0.015	0.772	0.017	0.005, 0.029	0.005

4.3.2.2 BMI allele scores to diet PCs results in ALSPAC

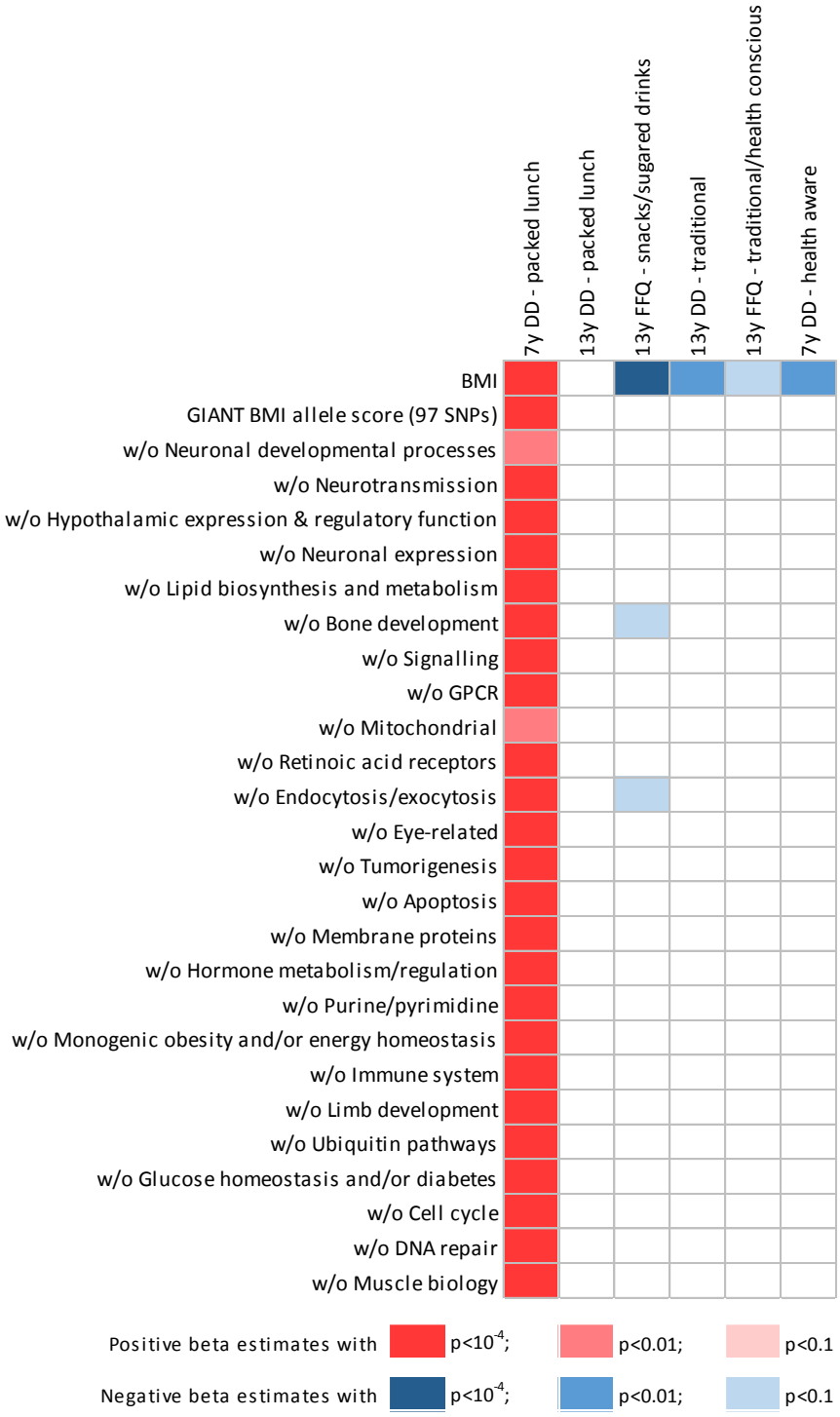
The 97 SNPs from Locke et al. were divided into overlapping subcategories based on the biological role of nearby genes, and LOO weighted allele scores were generated for each biological subcategory. Linear regression was performed to test the association between each of the 25 allele scores and the 6 diet PCs. Linear regression models were also fitted to test the association between the 97-SNP weighted allele score and the diet PCs. Models were adjusted for age and sex.

The heatmap in **Figure 11** shows the strength and direction of the relationships between the allele scores and the diet PCs.

Where there is suggestive evidence of associations between the LOO allele scores and diet PCs, the effect estimates of the allele scores on the diet PCs are directionally consistent with the effect estimates of BMI on the diet PCs. The age 7 diet diary "packed lunch" PC associated strongly with all the allele scores, though the association was weaker for the LOO neuronal development processes score and the LOO mitochondrial score. The age 13 FFQ "snacks/sugared drinks" PC displayed weak negative associations with the LOO bone development score and the LOO endocytosis/exocytosis allele score.

Figure 11 - Heatmap showing the strengths and effect directions of the relationships between the diet PCs and the BMI allele scores.

P-values and effect directions are from the model: diet PC ~ BMI allele score + age + sex. N=3,152-5,239. The first row of the heatmap shows the strengths and effect directions of the relationships between the diet PCs and BMI; p-values and effect directions are from the "diet ~ BMI + age + sex" model in **Table 12**.



4.4. Discussion

Using data from two major cohort studies (UK Biobank and ALSPAC), the analyses presented in this chapter sought to characterise the relationship between dietary behaviours and BMI. This involved both observational associations which highlighted links and the application of MR, where possible, to attribute direction of a causal pathway.

Observational analysis

Observational analyses identified several strong links between dietary behaviours and BMI. Associations observed in UK Biobank between macronutrient intake and BMI tended to be stronger than those observed in ALSPAC between the diet PCs and BMI, though this may, in part, be due to the far greater sample size available in UK Biobank.

Analyses in UK Biobank and ALSPAC both observed a negative association between adiposity and fibre intake, captured as g/day in UK Biobank and as a dietary pattern characterised by intake of high fibre foods such as fresh fruit and high fibre bread in ALSPAC. Previous studies have also observed negative associations between fibre intake and adiposity.[24,28]

Analysis of the relationship between the "snacks/sugared drink" intake PC (characterised by higher intakes of crisps, biscuits, chocolate, sweets, squash and fizzy drinks) and BMI in the ALSPAC teenagers at age 13 found that this dietary pattern was associated with a lower BMI, which is surprising. A possible explanation for this is that the teenagers are snacking rather than eating proper meals. Many of the foods in this dietary pattern are high in sugar, hence the negative association between this dietary pattern and BMI fits with the observed negative association between percentage sugar intake and BMI in UK Biobank.

Causal analysis

MR analyses were conducted to assess the causal effect of macronutrient intake on BMI. Macronutrients were instrumented using the SNPs identified in the UK Biobank diet

GWAS in Chapter 3, hence one-sample MR could not be conducted in UK Biobank as this would lead to overfitting.[210,211] Instead, two-sample MR was conducted using SNP-macronutrient coefficients from the UK Biobank GWAS and SNP-BMI coefficients from a large published BMI GWAS.[52]

MR, however, is limited by the availability of genetic instruments, which was problematic here since few genetic instruments exist for macronutrient intake, and none have been identified for the ALSPAC diet PCs. MR requires the instrument not to be associated with any confounders.[55] Three of the SNPs used as genetic instruments in the 2-sample MR diet to BMI analysis have been previously reported to be associated with several other traits (**Table 6**). Some of these reported traits may confound the relationship between diet and BMI, for example resting metabolic rate which was associated with rs838133 and rs516246.[151]

MR requires the outcome to only be associated with the instrument through the exposure.[55] However, this condition does not hold for the 97-SNP BMI score from Locke et al.[52] in MR analyses of BMI on diet since some of the loci are thought to influence appetite.[208,213] Instead, analyses were performed using LOO allele scores. The heatmaps of the associations in UK Biobank between the LOO allele scores and macronutrient intake show that the associations were weaker for allele scores without SNPs in the neuronal developmental processes, hypothalamic expression and regulatory function, and endocytosis/exocytosis categories. The association between the age 7 "packed lunch" PC and the LOO neuronal developmental processes allele score in ALSPAC was also weaker than other scores. This suggests that some of the BMI SNPs may influence dietary choices, and hence BMI.

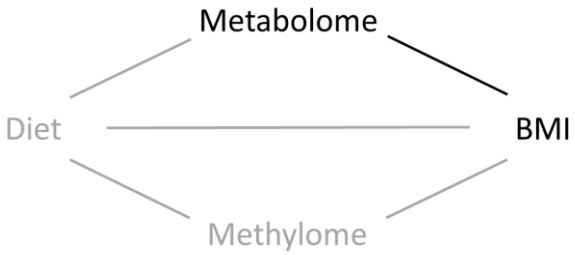
The evidence observed from the analyses undertaken in this chapter is not strong enough to come to a definitive conclusion on the causal relationship between dietary behaviour and BMI. Results from the causal analyses suggest that the relationship between diet and BMI is bidirectional. This is highly plausible since basal energy demands vary by BMI. The two-sample MR analysis results suggest that macronutrient intake (in particular, protein intake and polyunsaturated fat intake) may have a causal

effect on BMI. However, these results are based on single SNPs, and hence further investigation is needed for clarification. Results from analyses exploring the relationship between BMI allele scores and dietary behaviour imply that BMI also influences diet, since the LOO analyses suggest that the association between the BMI allele score and macronutrient intake is not fully explained by the BMI SNPs exerting an effect on BMI through dietary behaviour.

Future directions

Future analyses would benefit from the use of better dietary measures to identify robust genetic instruments for dietary behaviour that could be implemented in an MR framework (**3.4.3**). More refined measures of body composition, such as DXA assessments, could also be used since BMI alone does not give an adequate measure of health and wellbeing, as is illustrated by the association observed between a higher snacks/sugared drink intake and lower BMI.

CHAPTER 5. BMI AND THE METABOLOME



5.1. Introduction

Numerous studies have observed strong associations between adiposity and the human serum metabolome.[68-78] Most of these studies are cross-sectional and hence did not investigate the direction of causality. Some studies attempted to establish the direction of causality using genetic methods such as MR or by longitudinal analysis. Studies have used various adiposity measures, including BMI,[70,73,74] waist circumference,[71] android fat (%) and fat mass.[72] Two studies did an obese case-control analysis.[68,69] The following subsections review the literature describing the relationship between the human serum/plasma metabolome and adiposity.

5.1.1. Observational relationships between adiposity and the metabolome

5.1.1.1 Children

Wahl et al. and Perng et al. both used mass-spectrometry based approaches to compare the metabolite profiles of normal-weight children with those of obese children.[68,69] Wahl et al. compared serum metabolite profiles of normal-weight children (n=40) against obese children (n=80) and observed significant differences for 14 metabolite concentrations, including some amino acids (glutamine, methionine and proline) and phosphatidylcholines, and 69 metabolite ratios.[68] Compared to normal-weight children, amino acid concentrations in obese children were 20%, 22% and 30% lower for glutamine, methionine and proline respectively.

Rather than studying individual metabolites, Perng et al. used PCA to consolidate 345 metabolites (from plasma) into 18 factors and then compared the factor scores of 'lean' children (n=150) and 'obese' children (n=84).[69] These metabolite factors capture patterns based on correlations between the individual metabolites. Perng et al. observed significant differences for two of the factors: one characterised by positive loadings of amino acids phenylalanine, valine, leucine and isoleucine, and the other by positive loadings of androgen hormones. Both factors were higher in obese children

than lean children. The advantages of using PCA in this context are that it reduces the total number of variables to be studied in relation to adiposity, and that it enables identification of the metabolite patterns that are most strongly associated with adiposity. However, the use of PCA here results in a loss of information compared to using the individual metabolites.

5.1.1.2 Young adults

Studies of young adults have also observed strong links between adiposity and the metabolome including lipoproteins, amino acids and fatty acids.[70,71] Würtz et al. studied the relationship between BMI and the serum metabolite profile in >12,000 adolescents and young adults from four population-based cohorts in Finland.[70] They used an NMR platform to quantify 67 serum metabolic measures. They also assayed 15 plasma metabolic measures including inflammatory markers and hormones. Würtz et al. observed cross-sectional associations between BMI and 68 of the 82 metabolites they studied. These included positive associations with phenylalanine, valine, leucine and isoleucine, consistent with findings by Perng et al.[69] Würtz found that elevated BMI was associated with adverse changes in the metabolite profile.

Bogl et al., in their study of young adults (n=1368, mean age 24.3 ± 0.1 years), used an NMR platform to quantify serum metabolites and found that waist circumference (WC) was associated with 50 of the 56 metabolites they studied.[71] Their findings included positive correlations with VLDLs, LDLs, small HDLs, total cholesterol and ApoB, and negative correlations with large HDLs. Their findings also included associations with some fatty acids, amino acids and glycolysis-related metabolites. They did not observe associations with lactate, histidine, acetate or creatinine. Their findings are mostly consistent with those by Würtz et al.,[70] though for some metabolites one of the studies detected an association where the other did not. In a smaller sample (n=286, mean age 28.7 ± 0.2 years), they studied other obesity measures, including android fat (%) and subcutaneous fat, and found that abdominal fat is overall most strongly associated with an adverse metabolite profile.

5.1.1.3 Middle-aged and older adults

There have also been studies of (or including) middle-aged and older adults.[72-74] Boulet et al. studied women between the ages of 37 and 59 years.[72] Their analyses are detailed, however their sample size is quite small (59 adults). They used mass-spectrometry to quantify 138 plasma metabolites and then performed PCA to reduce the dimensionality of the metabolite data. They found that obese women had significantly higher branched-chain amino acid levels than lean or overweight women; this is consistent with findings described above in children and young adults.[69-71]

Ho et al. also observed associations between BMI and several metabolites including positive associations with branched-chain amino acids.[73] They studied 2,383 adults (mean age 55 ± 10 years) from the Framingham Offspring cohort and used mass spectrometry to measure levels of 217 plasma metabolites, of which 69 were associated with BMI.

Moore et al. performed a meta-analysis of three studies examining the cross-sectional associations between metabolite levels and BMI and identified 37 metabolites associated with BMI.[74] Consistent with Würtz et al.,[70] they observed positive associations with lactate, glycerol and the amino acids valine, tyrosine, leucine and phenylalanine. The studies (in the Moore et al. meta-analysis) used mass spectrometry to measure levels of 317 metabolites from blood samples.

5.1.2. Causality in the relationship between adiposity and metabolites

Würtz et al. found that elevated BMI was associated with adverse changes in the metabolite profile in young adults.[70] Having made this observation, they then used MR to investigate causality.[54] Their genetic-predisposition score for elevated BMI was associated with multiple metabolites, which suggests that adiposity has a causal effect on the metabolic profile. They also followed-up c.1,500 young adults over a 6-year period and found that the change in BMI was associated with changes in the metabolite profile, suggesting that the metabolite profile is modifiable.

Whereas Würtz et al. studied a detailed metabolite profile made up of more than 80 metabolites, Holmes et al. and Freathy et al. focused on a small number of metabolic traits.[75,76] Holmes et al. conducted MR analyses to investigate the effect of BMI on 14 cardiometabolic traits in adults.[75] Their results suggest that BMI has a causal effect on some of these traits including fasting glucose and HDL cholesterol, but not LDL cholesterol where the instrumental variable (IV) analysis estimate was directionally opposite to the observational estimate. The observational effect estimate of BMI on HDL cholesterol was -0.02 (-0.02, -0.02) mmol/l per 1kg/m^2 increase in BMI, and the IV estimate of BMI on HDL cholesterol was -0.02 (-0.03, -0.01) mmol/l per 1kg/m^2 increase in BMI.

Freathy et al. tested the association between genetic variants in *FTO* and 10 metabolic traits in adults and observed associations for 4 of them: triglycerides, HDL cholesterol, fasting glucose and fasting insulin.[76] They used a triangulation approach[76] to investigate causality and found that the effect sizes for the pairwise associations between *FTO*, BMI and the metabolic traits are consistent with BMI having a causal effect on the metabolic traits. The observational and IV estimates for the effect of BMI on HDL cholesterol were both negative, which is consistent with the findings of Würtz et al. and Holmes et al.[70,75]

5.1.3. Aims and objectives

Whilst the work of Würtz et al. has demonstrated the effect of adiposity on metabolic signatures from early adulthood onwards, there is no published data exploring the relationship between adiposity and the metabolome in childhood. The growing obesity epidemic impacting across the lifecourse is a major public health concern in many countries, therefore understanding the causes and consequences of adiposity is important at all ages. The aim of these analyses is to explore the relationship between adiposity and the metabolome during childhood and adolescence and to determine whether the relationships during this crucial developmental period are consistent with or differ from those in adulthood. Adiposity is measured by BMI.

Chapter objectives:

i) Assess the cross-sectional relationship between BMI and the metabolome in the ALSPAC participants at ages 7 and 15 years
ii) Conduct MR analyses to explore the causal effect of BMI on the metabolome at age 7 years
iii) Conduct longitudinal analyses to test whether change in BMI between age 7 and 15 is associated with changes in the metabolome between age 7 and 15

5.2. Methods

5.2.1. Metabolite quantification

Metabolite profiles for ALSPAC participants have been generated from serum samples taken at age 7 and 15 years.[110] Samples from the 7-year-olds are non-fasting samples, whereas samples from the 15-year-olds are fasting samples. The metabolic measures were quantified using a high-throughput serum nuclear magnetic resonance (NMR) platform, which measures levels of >200 metabolites.[57,110]

5.2.2. Data preparation

5.2.2.1 7-year-olds' metabolites

Metabolites with skewed distributions (skewness > 2) were normalized by applying a log-transformation prior to analysis, like in the analyses by Würtz et al.[214] All metabolite concentrations were scaled to standard deviation units, allowing for easy comparison of effect sizes in later analyses.

PCA could be used to generate metabolite patterns from the individual metabolites. This would reduce the number of variables to analyse, however it would also result in a loss of information. The decision of whether to use metabolite patterns or individual metabolites depends on the question of interest, so for these analyses individual metabolite data was used. A major aim of these analyses is to compare the effect sizes of BMI on metabolites in childhood and adolescence with previously observed effect

sizes from the literature, and principal components would not be comparable across studies.

5.2.2.2 15-year-olds' metabolites

In order to be able to compare metabolite data from age 15 with that from age 7, similar transformations were applied to the 15-year-olds' metabolite data. Specifically, metabolite measures in the 15-year-olds were log-transformed if (and only if) that metabolite measure had a skewed distribution in the 7-year-olds. The 15-year-olds' metabolite data was then scaled to 15-year-olds' metabolite standard deviation units. Although this means that the 15-year-olds' metabolite data does not have the same scale as the 7-year-olds, scaling the data to standard deviation units from the 15-year-olds rather than the 7-year-olds was more appropriate because the metabolite distributions at age 15 are different to those at age 7. Scaling the data this way enables detection of differences in metabolite profiles between individuals, rather than typical within-individual changes in metabolite levels as the children grow older.

5.2.2.3 Correction for multiple testing

Since most of the metabolic measures are strongly correlated, use of a Bonferroni correction would be overly conservative. A similar approach to Würtz et al. was taken: principal components analysis (PCA) was performed to consolidate the metabolic measures and calculate how many principal components (PCs) are needed to account for at least 95% of the variance.[70] It was calculated that for the 7-year-olds' metabolite data the top 18 PCs are needed to account for 95% of the variance. Since cross-sectional, longitudinal and MR analyses are being performed, 3×18=54 was taken to be the number of tests for the Bonferroni correction, which resulted in defining $p<0.05/54\approx0.001$ to be the statistical significance threshold, based on the arbitrary p-value threshold of $p<0.05$. PCA was performed in R (version 3.3.3) using the *pca* function from the *pcaMethods* package.

5.2.3. Cross-sectional analyses

Cross-sectional analyses were performed at ages 7 and 15 years. First, the cross-sectional relationship between BMI and the metabolome at age 7 was investigated. For each metabolic measure, linear regression was performed (using the *lm* function from the *stats* package), with BMI as the explanatory variable and the metabolic measure as the outcome variable, adjusting for age and sex.

Metabolite ~ BMI + age + sex

(ALSPAC children at age 7 years)

To be able to compare these results from the 7-year-olds with the Würtz et al. young adults' results (mean age 26 years) and results from the ALSPAC children at age 15, the analyses were repeated at 15 years for the metabolites for which strong associations had been observed at age 7 years and for which cross-sectional results were available from the Würtz et al. young adults.[70]

Metabolite ~ BMI + age + sex

(ALSPAC children at age 15 years)

Since Würtz et al. applied log-transformations to their metabolites according to the skewness of their data, the metabolites that they transformed differ a little to those that are transformed in these analyses. To be consistent with these analysis models, log-transformations or inverse log-transformations were applied to their results according to the log-transformations that were applied to the ALSPAC metabolite variables.

5.2.4. Mendelian randomization analyses

Having observed cross-sectional associations between BMI and several of the metabolite measures, MR analyses were performed in R (version 3.3.3) using the two-stage least squares method (using the *ivreg* function from the *AER* package) to assess whether the observed cross-sectional associations between BMI and metabolites at age 7 years

represent a causal effect of BMI on the metabolome. The GIANT BMI score was used as the genetic instrument for BMI (**2.2.4**).[52] MR-Egger regression was performed (using the *mr_egger* function from the *MendelianRandomization* package) (**2.2.4.2**), which is a pleiotropy-robust method, to assess the validity of using the GIANT BMI score as a genetic instrument in these analyses.

5.2.5. Longitudinal analyses

Longitudinal analyses were performed to investigate whether change in BMI between the ages of 7 and 15 years is associated with the change in metabolite levels across the same period.

One approach could be to use the raw changes in BMI and metabolites between the ages of 7 and 15 years. That is, defining change in BMI as the difference in kg/m^2 between their absolute BMI at age 7 and age 15. However, using the raw change in BMI between these ages may not be a good representation of change in adiposity because of the normal physical development a child undergoes throughout childhood and adolescence. For example, a child with BMI of $20kg/m^2$ at age 7 years is usually considered to be overweight, but a 15-year-old with the same BMI is a healthy weight.

Instead, to gain a better understanding of how a child's BMI changes over time compared to their contemporaries, z-scores were used. BMI z-scores were generated in the 7-year-olds and 15-year-olds separately and then for each child the difference between their two z-scores was calculated. For example, if a child's BMI is much lower than average at age 7 but by the age of 15 their BMI is only slightly below average then they will have a positive increase in z-score. The change in z-scores for the metabolites was also calculated.

For each of the metabolites, linear regression was performed (using the *lm* function from the *stats* package), with change in metabolite z-score between 7 and 15 years as the outcome variable and change in BMI z-score between 7 and 15 years as the explanatory variable, adjusting for age at baseline (7 years) and sex.

(Change in metabolite z-score) ~ (Change in BMI z-score) + (age at baseline) + sex

5.3. Results

5.3.1. Cross-sectional analyses

At age 7 years associations were observed between BMI and 108 of the 160 metabolite measures studied, using p-value threshold $p < 0.001$ (**Figure 12**). Associations were observed for all but one of the VLDL concentration measures, the majority of the HDL concentration measures, and several other measures including apolipoproteins, fatty acids and amino acids. The VLDLs were positively associated with BMI and the HDLs were negatively associated with BMI.

The forest plot in **Figure 13** shows how the age 7 results compare with the age 15 results and the Würtz young adult results (plot only includes metabolites that were associated at age 7 and have results available at age 15 and in Würtz young adults). The Würtz metabolites were log-transformed so that they are on the same scale as the ALSPAC metabolites. The effect estimates are mostly in the same direction in all three samples, the exceptions being for estimated fatty acid chain length (effect estimates are positive at age 7 and 15 and negative in the young adults) and lactate (effect estimates are negative at age 7 and positive at age 15 and in the young adults).

If the magnitude of the effect of BMI on metabolites changes with age and this change is consistent from childhood into young adulthood, then one would expect the age 15 effect estimates to lie between (or at least overlap with both of) the age 7 and young adult effect estimates. This is most clearly observed for creatinine where the age 15 effect estimate lies between the age 7 and young adult effect estimates and the confidence intervals do not overlap.

5.3.2. MR analyses

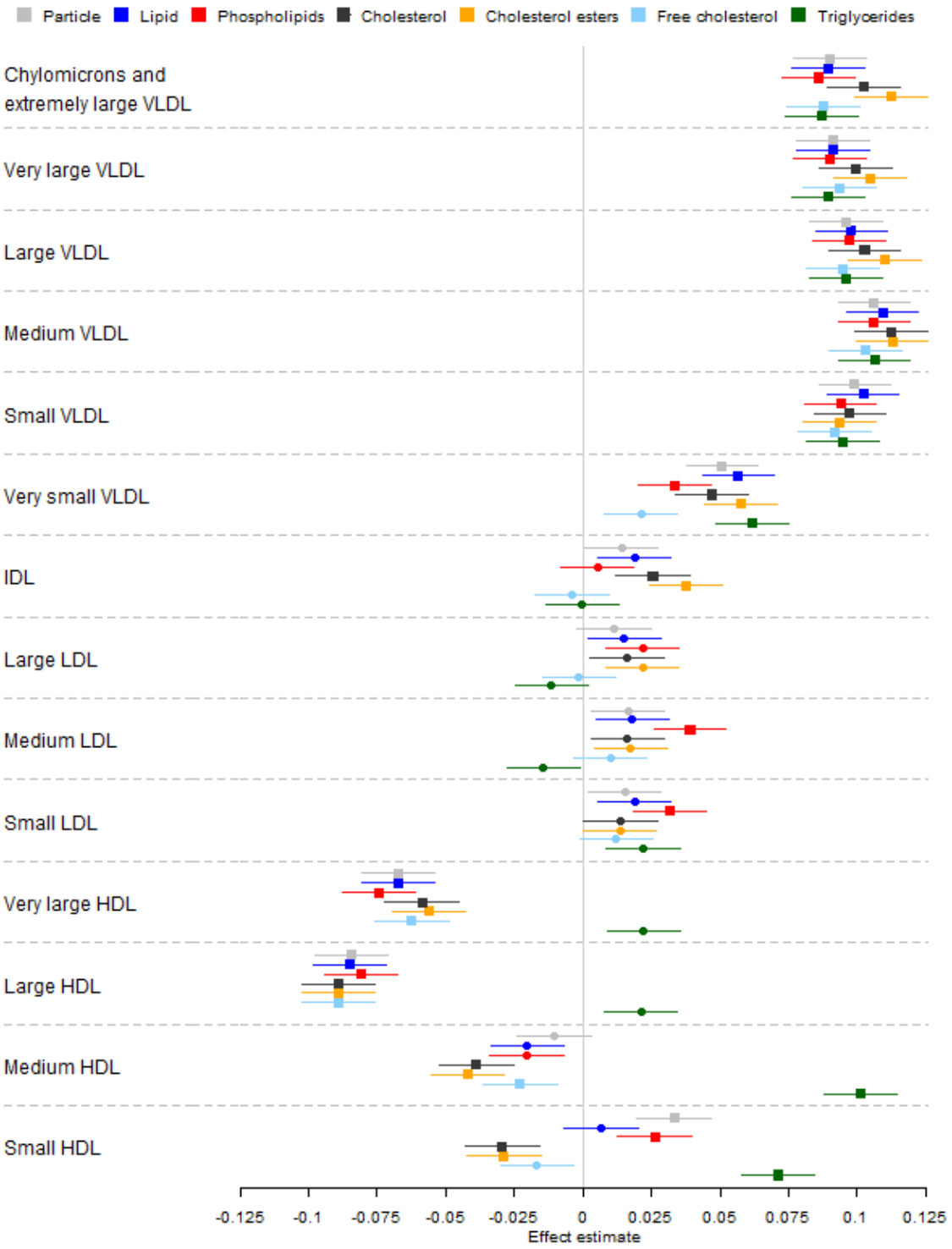
The causal effect estimates from MR analyses were mostly directionally consistent with the observational estimates, however the confidence intervals of the causal effect estimates were wide and, for 77 (out of 106) metabolites, spanned zero.

For most of the metabolites the cross-sectional effect estimate confidence interval falls completely within the MR effect estimate confidence interval. The only exceptions are three amino acids (histidine, valine and tyrosine) though there is still some overlap. The MR effect estimates for valine and tyrosine are directionally consistent with but larger than the cross-sectional effect estimates. The cross-sectional and MR effect estimates for histidine are directionally inconsistent – the cross-sectional estimate is positive whereas the MR estimate is negative (though its confidence interval spans zero).

Metabolites with MR confidence intervals that did not span zero include: cholesterol and cholesterol esters for most of the VLDL sub-classes; all the particle types for the medium and small VLDL sub-classes; HDL triglycerides; some amino acids; creatinine; and glycoprotein acetyls.

Figure 12 – Forest plots of cross-sectional associations of metabolites and BMI in the ALSPAC children at age 7.

Effect estimates are the 1-SD increase in metabolite concentration per unit increase in BMI (kg/m^2). The point estimates are represented by a square p<0.001, and a circle otherwise. The lines though the points are the 95% CIs. The metabolites have been divided into two plots. The first plot (below) shows the results for the fourteen lipoprotein subclasses and their lipid measures, and the second plot (following page) shows results for the remaining metabolite measures. N=5,414.



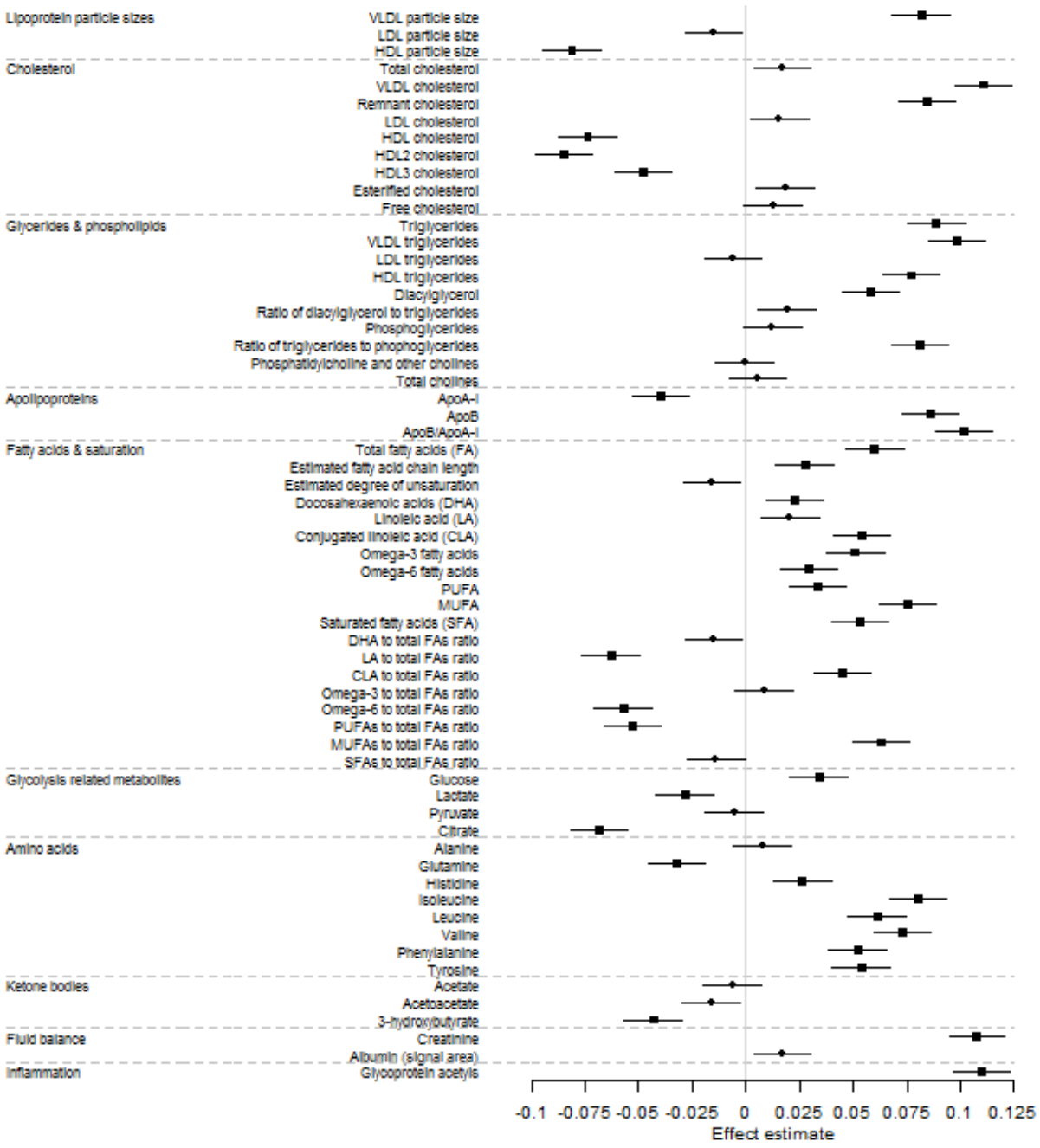

Lipoprotein particle sizes
VLDL particle size
LDL particle size
HDL particle size
Cholesterol
Total cholesterol
VLDL cholesterol
Remnant cholesterol
LDL cholesterol
HDL cholesterol
HDL2 cholesterol
HDL3 cholesterol
Esterified cholesterol
Free cholesterol
Glycerides & phospholipids
Triglycerides
VLDL triglycerides
LDL triglycerides
HDL triglycerides
Diacylglycerol
Ratio of diacylglycerol to triglycerides
Phosphoglycerides
Ratio of triglycerides to phosphoglycerides
Phosphatidylcholine and other cholines
Total cholines
Apolipoproteins
ApoA-I
ApoB
ApoB/ApoA-I
Fatty acids & saturation
Total fatty acids (FA)
Estimated fatty acid chain length
Estimated degree of unsaturation
Docosahexaenoic acids (DHA)
Linoleic acid (LA)
Conjugated linoleic acid (CLA)
Omega-3 fatty acids
Omega-6 fatty acids
PUFA
MUFA
Saturated fatty acids (SFA)
DHA to total FAs ratio
LA to total FAs ratio
CLA to total FAs ratio
Omega-3 to total FAs ratio
Omega-6 to total FAs ratio
PUFAs to total FAs ratio
MUFAs to total FAs ratio
SFAs to total FAs ratio
Glycolysis related metabolites
Glucose
Lactate
Pyruvate
Citrate
Amino acids
Alanine
Glutamine
Histidine
Isoleucine
Leucine
Valine
Phenylalanine
Tyrosine
Ketone bodies
Acetate
Acetoacetate
3-hydroxybutyrate
Fluid balance
Creatinine
Albumin (signal area)
Inflammation
Glycoprotein acetyls
-0.1 -0.075 -0.05 -0.025 0 0.025 0.05 0.075 0.1 0.125
Effect estimate

Figure 13 – Forest plot comparing cross-sectional effect estimates from the ALSPAC children at ages 7 and 15 and the Würtz young adults.

Effect estimates are the 1-SD increase in metabolite concentration per unit increase in BMI (kg/m^2). The lines though the points are the 95% CIs. N=5,305-5,416 at age 7. N=3,094-3,286 at age 15. N=12,664 in adults.

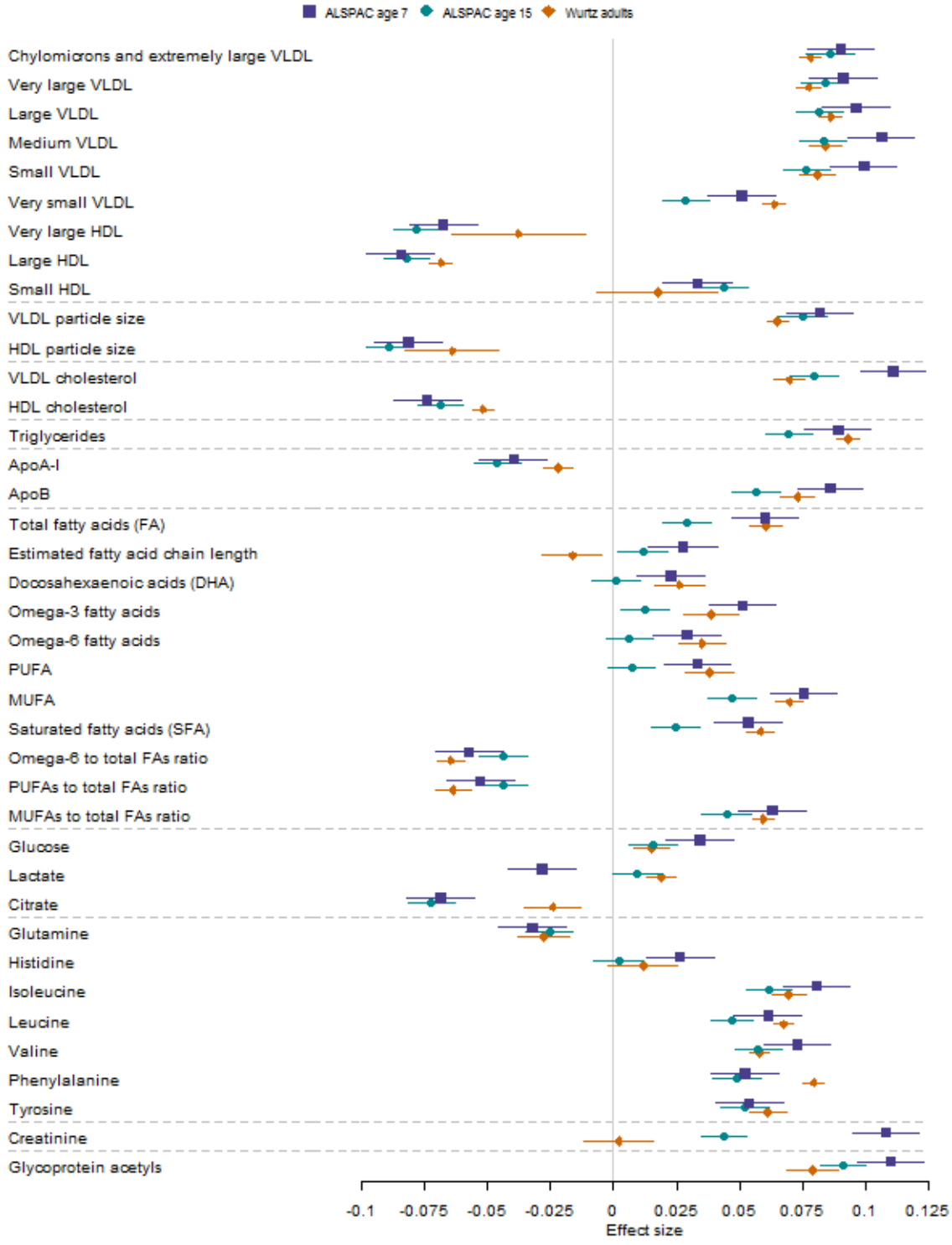


Figure 14 shows a correlation plot of the cross-sectional and MR analyses, showing the line of best fit (gradient=0.84; intercept=0.01), the effect estimates and confidence intervals, and $R^2 = 0.84$. Hence overall the cross-sectional and MR effect estimates are similar.

Figure 15 shows forest plots comparing MR and cross-sectional results.

For each metabolite, a classic z-statistic was calculated to compare the cross-sectional and MR effect estimates. Evidence was not observed for differences between the estimates – out of the 106 metabolites the smallest p-values were p=0.052 (histidine) and p=0.087 (valine). Although some of the beta estimates differ a lot between the cross-sectional and MR analyses, the MR confidence intervals are wide which makes it hard to rule out or confirm a causal effect of BMI on metabolites.

The results of the MR-Egger sensitivity analyses (in appendix) suggest that pleiotropy may be an issue for the MR analyses of BMI with tyrosine and creatinine since the MR-Egger intercepts have small p-values.

Overall, these MR analyses give evidence of a causal effect of BMI in childhood on several metabolites, including VLDLs, HDL triglycerides, amino acids and creatinine.

Figure 14 – Correlation plot of effect estimates (and 95% CIs) from cross-sectional and MR analyses.

The effect estimates are 1-SD increase in metabolite concentration per unit increase in BMI (kg/m^2). The data points are the effect estimates, and the error bars are the 95% CIs. The diagonal grey dotted line is the y=x line, and the red dotted line is the line of best fit.

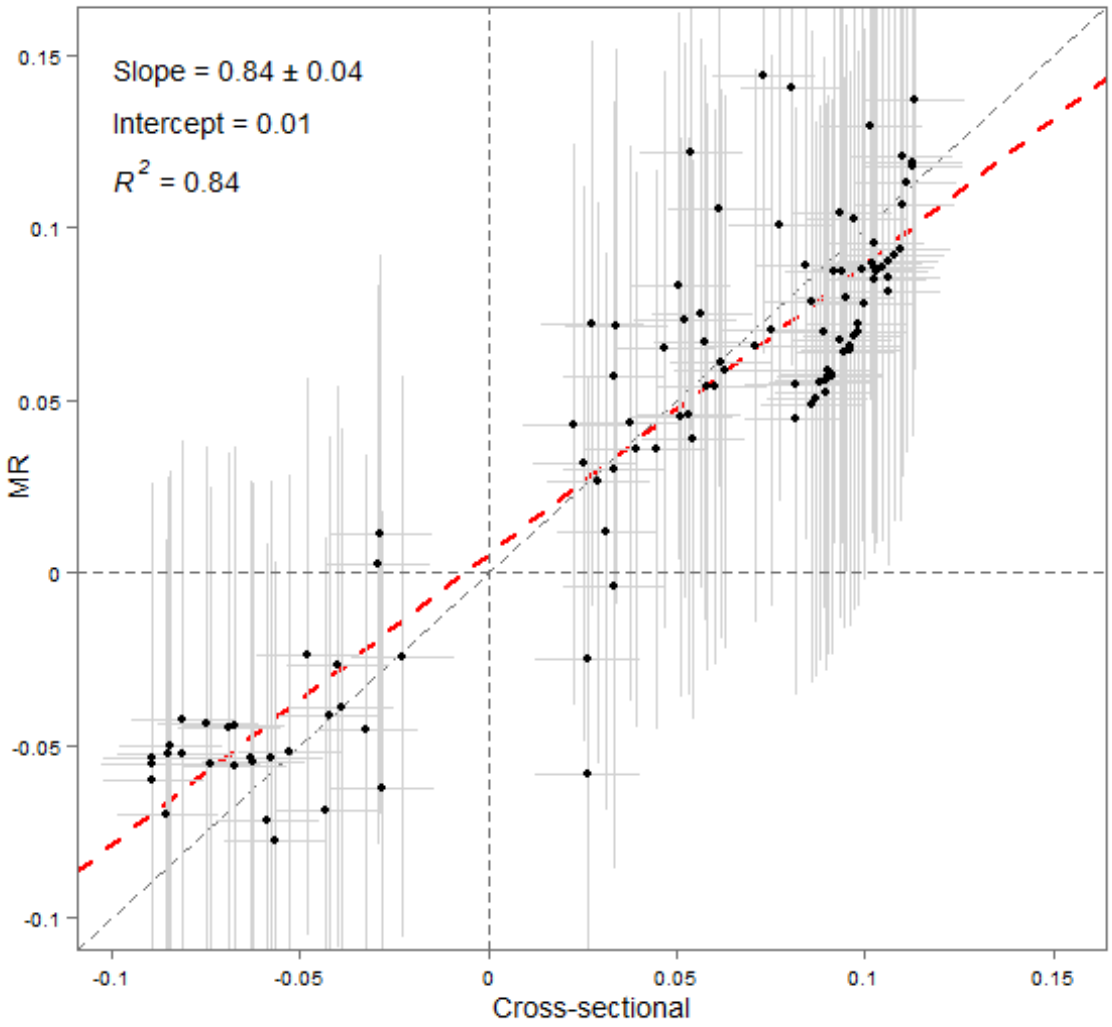
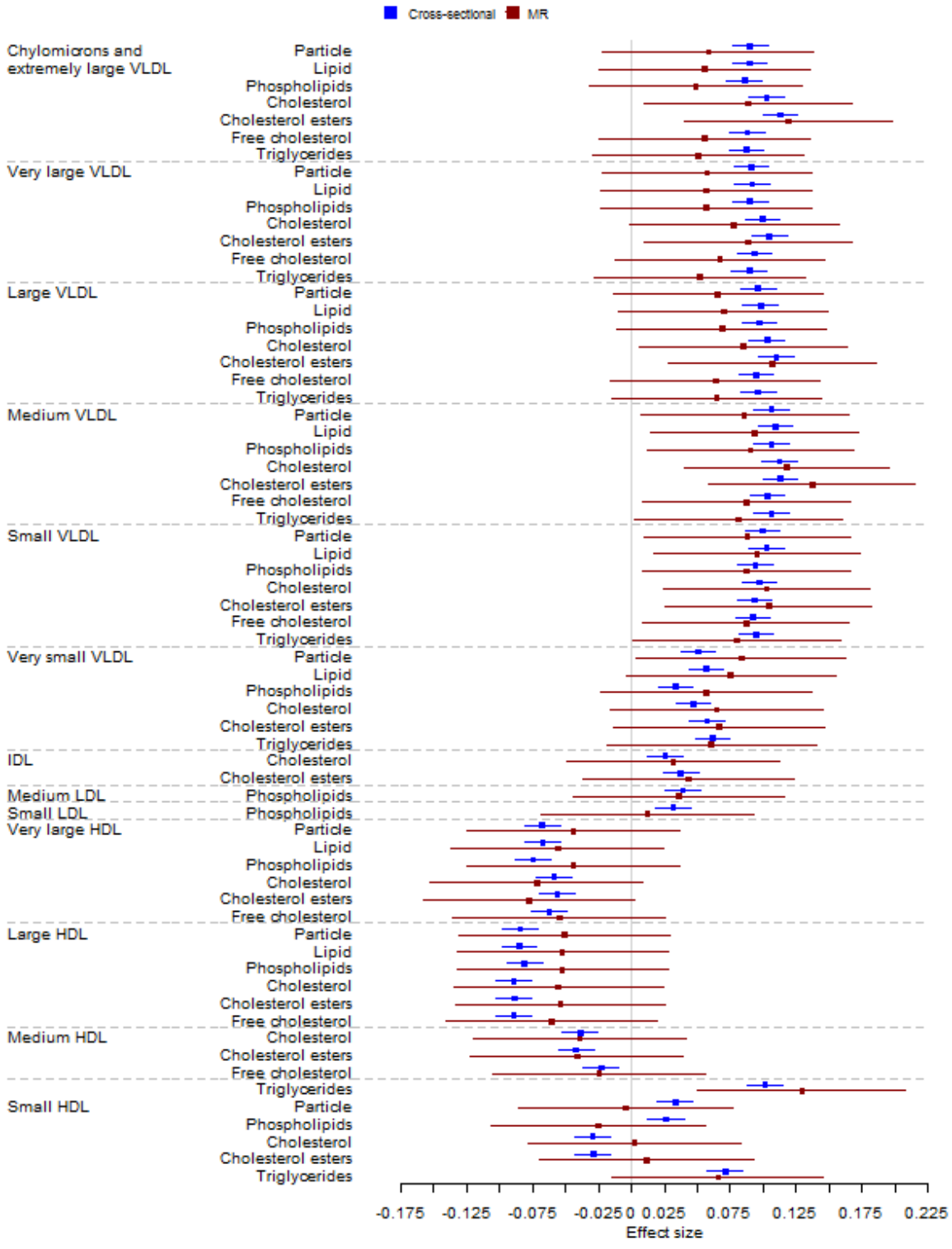
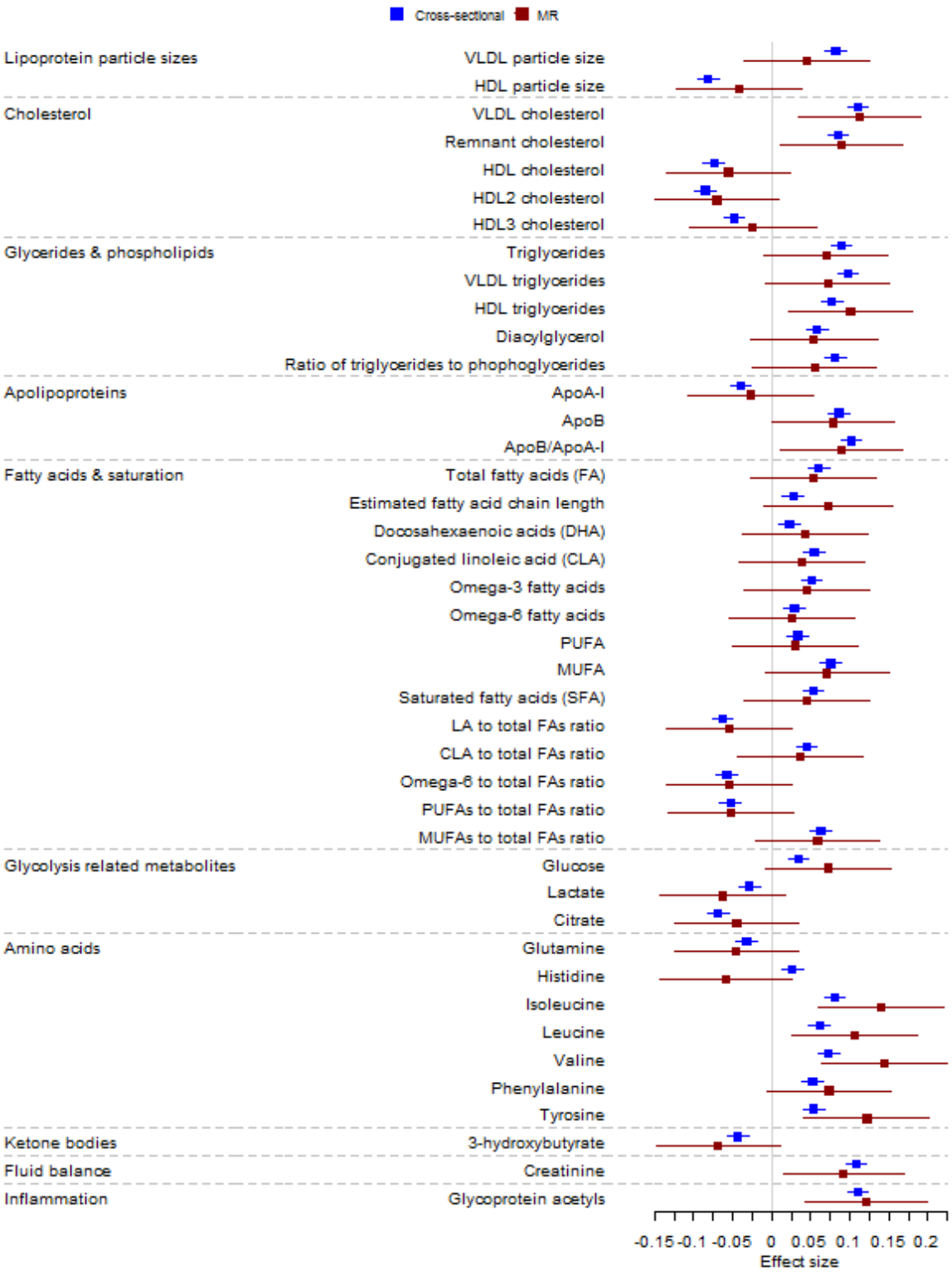


Figure 15 – Forest plots comparing effect estimates from cross-sectional and MR analyses in the ALSPAC children at age 7.

Effect estimates are the 1-SD increase in metabolite concentration per unit increase in BMI (kg/m^2). The lines though the points are the 95% CIs. Metabolites are only included if they were cross-sectionally associated with BMI at age 7 years. The metabolites have been divided into two plots. The first plot (below) shows the results for the fourteen lipoprotein subclasses and their lipid measures, and the second plot (following page) shows results for the remaining metabolite measures. N=5,305-5,416 for cross-sectional analyses. N=4,380-4,469 for MR analyses.




Cross-sectional
MR
Lipoprotein particle sizes
VLDL particle size
HDL particle size
Cholesterol
VLDL cholesterol
Remnant cholesterol
HDL cholesterol
HDL2 cholesterol
HDL3 cholesterol
Glycerides & phospholipids
Triglycerides
VLDL triglycerides
HDL triglycerides
Diacylglycerol
Ratio of triglycerides to phophoglycerides
Apolipoproteins
ApoA-I
ApoB
ApoB/ApoA-I
Fatty acids & saturation
Total fatty acids (FA)
Estimated fatty acid chain length
Docosahexaenoic acids (DHA)
Conjugated linoleic acid (CLA)
Omega-3 fatty acids
Omega-6 fatty acids
PUFA
MUFA
Saturated fatty acids (SFA)
LA to total FAs ratio
CLA to total FAs ratio
Omega-6 to total FAs ratio
PUFAs to total FAs ratio
MUFAs to total FAs ratio
Glycolysis related metabolites
Glucose
Lactate
Citrate
Amino acids
Glutamine
Histidine
Isoleucine
Leucine
Valine
Phenylalanine
Tyrosine
Ketone bodies
3-hydroxybutyrate
Fluid balance
Creatinine
Inflammation
Glycoprotein acetyls
-0.15 -0.1 -0.05 0 0.05 0.1 0.15 0.2
Effect size

5.3.3. Longitudinal analyses

All the VLDL measures show positive longitudinal associations between change in level and change in BMI. Phospholipids in LDLs also show positive associations. Most of the very large and large HDL measures show negative longitudinal associations, except for the triglycerides. Overall, the triglyceride effect estimates differ most from the other particles in the LDL and HDL categories, and the triglyceride effect estimates tend to be in the opposite direction.

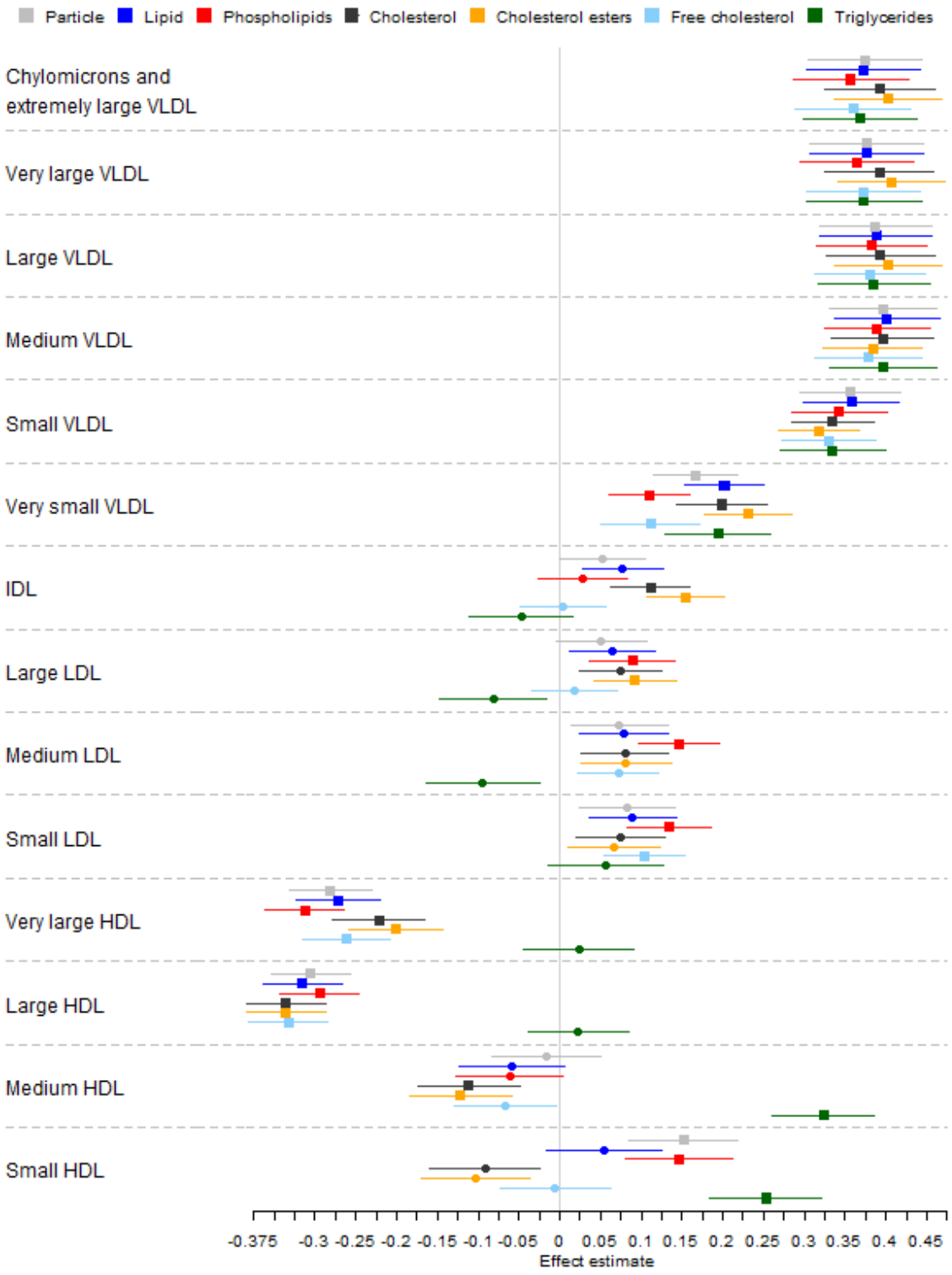
Change in VLDL particle size is positively associated with change in BMI. LDL and HDL particle sizes are negatively associated with BMI. In the cholesterol category of metabolite measures, change in VLDL and remnant cholesterol are positively associated with change in BMI, and change in HDL cholesterol is negatively associated. In the glycerides and phospholipids category, triglycerides (except LDL triglycerides) are positively longitudinally associated with BMI.

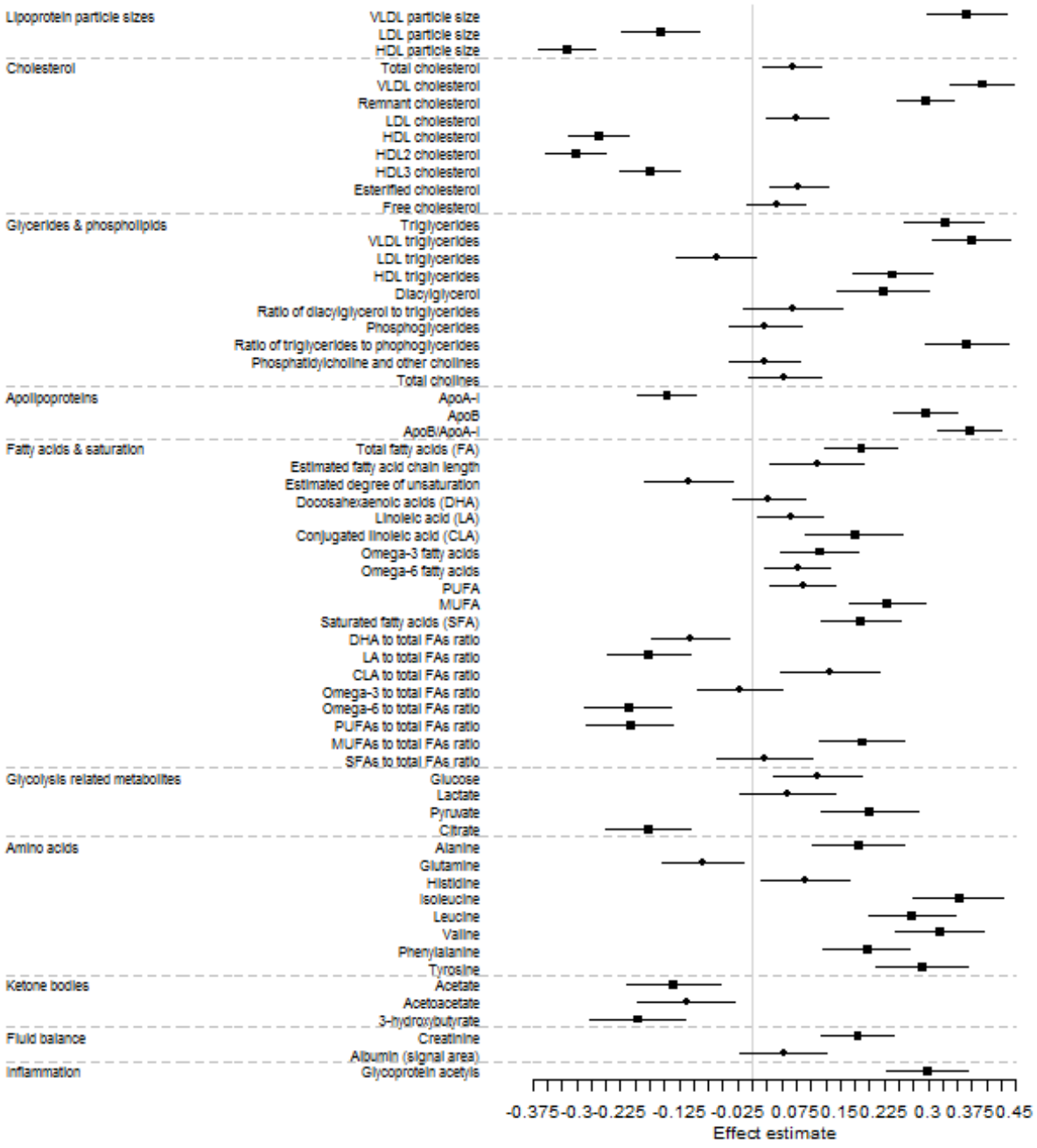
Change in ApoA-I is negatively associated with change in BMI, whereas change in ApoB is positively associated. Change in some fatty acids, including MUFA, are associated with change in BMI. Most of the amino acids are positively longitudinally associated with BMI, as are some other metabolites including pyruvate, creatinine and glycoprotein acetyls.

Figure 16 shows forest plots of the longitudinal results.

Figure 16 – Forest plots showing effect estimates for the relationship between change in metabolite z-score and change in BMI z-score between the age 7 and 15 years.

Effect estimates are the 1-SD increase in metabolite concentration per unit increase in BMI z-score. The point estimates are represented by a square p<0.001, and a circle otherwise. The lines though the points are the 95% CIs. The metabolites have been divided into two plots. The first plot (below) shows the results for the fourteen lipoprotein subclasses and their lipid measures, and the second plot (following page) shows results for the remaining metabolite measures. N=2,049-2,181.




Lipoprotein particle sizes
VLDL particle size
LDL particle size
HDL particle size
Cholesterol
Total cholesterol
VLDL cholesterol
Remnant cholesterol
LDL cholesterol
HDL cholesterol
HDL2 cholesterol
HDL3 cholesterol
Esterified cholesterol
Free cholesterol
Glycerides & phospholipids
Triglycerides
VLDL triglycerides
LDL triglycerides
HDL triglycerides
Diacylglycerol
Ratio of diacylglycerol to triglycerides
Phosphoglycerides
Ratio of triglycerides to phosphoglycerides
Phosphatidylcholine and other cholines
Total cholines
Apolipoproteins
ApoA-I
ApoB
ApoB/ApoA-I
Fatty acids & saturation
Total fatty acids (FA)
Estimated fatty acid chain length
Estimated degree of unsaturation
Docosahexaenoic acids (DHA)
Linoleic acid (LA)
Conjugated linoleic acid (CLA)
Omega-3 fatty acids
Omega-6 fatty acids
PUFA
MUFA
Saturated fatty acids (SFA)
DHA to total FAs ratio
LA to total FAs ratio
CLA to total FAs ratio
Omega-3 to total FAs ratio
Omega-6 to total FAs ratio
PUFAs to total FAs ratio
MUFAs to total FAs ratio
SFAs to total FAs ratio
Glycolysis related metabolites
Glucose
Lactate
Pyruvate
Citrate
Amino acids
Alanine
Glutamine
Histidine
Isoleucine
Leucine
Valine
Phenylalanine
Tyrosine
Ketone bodies
Acetate
Acetoacetate
3-hydroxybutyrate
Fluid balance
Creatinine
Albumin (signal area)
Inflammation
Glycoprotein acetyls
-0.375 -0.3 -0.225 -0.125 -0.025 0.075 0.15 0.225 0.3 0.375 0.45
Effect estimate

5.4. Discussion

These analyses show strong evidence of associations between BMI and several metabolite measures in childhood and adolescence. The results of the MR analyses indicate that a genetic predisposition to higher BMI is associated with metabolite levels in childhood, which suggests that adiposity has a causal effect on the childhood metabolic profile, however for several of the metabolites the evidence is not conclusive.

Cross-sectional findings

Overall, these cross-sectional results in childhood and adolescence are similar to previously published results in adult population samples by Würtz et al. and Bogl et al.[70,71] Consistent with those studies, BMI is positively associated with VLDLs and triglycerides and negatively associated with large HDLs. Also consistent with those studies, BMI is positively associated with ApoB, the MUFAs to total FAs ratio, isoleucine, leucine, valine, phenylalanine, tyrosine and glycoprotein; and negatively associated with HDL cholesterol, the omega-6 to total FAs ratio, the PUFAs to total FAs ratio and glutamine.

The cross-sectional effect estimates from the ALSPAC 7- and 15-year-olds are mostly directionally consistent with results by Moore et al.[74] However the ALSPAC 7-year-olds' cross-sectional effect estimates are directionally inconsistent with Moore et al. for lactate and histidine. The cross-sectional effect estimate for lactate is negative in the ALSPAC 7-year-olds, but positive in the ALSPAC 15-year-olds, the Würtz young adults, the Moore adults, and also the Ho et al. adults.[73] The cross-sectional effect estimate for histidine is negative in the Moore adults, but positive in the ALSPAC 7- and 15-year-olds and the Würtz young adults (though the CIs for the 15-year-olds and the Würtz young adults span zero).

Figure 13 shows the cross-sectional effect estimates for the relationships between BMI and metabolome observed in the ALSPAC 7-year-olds, the ALSPAC 15-year-olds, and the Würtz young adults. For most of the metabolites, the cross-sectional effect sizes are similar between the groups and the confidence intervals overlap. This suggests that the

cross-sectional associations between BMI and these metabolites are consistent in direction and magnitude between from childhood, throughout adolescence and into young adulthood. The most notable exceptions to this are for estimated fatty acid chain length, lactate and creatinine. In all three of these exceptions, the age 15 cross-sectional effect estimate lies between the age 7 cross-sectional estimate and the young adult cross-sectional estimate, which suggests that the cross-sectional relationship between BMI and these metabolites changes with age.

Serum creatinine is a commonly used index of renal function due to its inverse relationship with glomerular filtration rate (GFR).[215] Glomerular filtration is the process by which the kidneys clear excess waste products (such as creatinine and urea) and fluids from the blood. Other factors known to be associated with GFR include age, sex, ethnicity, albumin concentration and urea nitrogen concentration.[216] Since GFR is lower in older adults, one could expect serum creatinine concentration to be higher in older adults. Obesity, along with diabetes and hypertension, is a known risk factor for chronic kidney disease.[217] Therefore, it is reasonable to expect creatinine levels to be higher in people with higher BMIs, which is consistent with the cross-sectional findings in the ALSPAC children (**Figure 13**).

These results show that the cross-sectional effect magnitude of BMI on creatinine decreases with age (**Figure 13**). A possible explanation for this could be that since kidney function tends to decrease with age anyway, the effect of BMI on kidney function becomes less pronounced with age.

Mendelian randomization findings in the 7-year-olds

The results of the MR analyses suggest that adiposity may have a causal effect on the metabolite levels in childhood, however for several of the metabolites the evidence is not conclusive since confidence intervals are wide and several span zero. The observational and MR estimates for the effect of BMI on HDL cholesterol were both negative, consistent with previous findings.[70,75,76] Strong evidence of causal effects was observed for isoleucine, leucine, valine and tyrosine both in these analyses and in analyses by Würtz et al.[70]

The 7-year-olds' causal effect estimates and observational effect estimates were mostly directionally consistent, or if not then the observational point estimate was contained within the causal estimate's confidence interval. The exception to this was for histidine, where the observational effect estimate was positive and the causal effect estimate was negative. These conflicting results for histidine are not surprising, since the cross-sectional association with BMI was not observed in the ALSPAC 15-year-olds, and since Moore et al. observed a negative cross-sectional effect estimate in adults.[74]

Longitudinal results

The longitudinal results are broadly consistent with results from longitudinal analyses by Würtz et al., including positive associations of BMI with VLDL, VLDL cholesterol, total fatty acids, omega-3 fatty acids, MUFA, saturated fatty acids and branched-chain and aromatic amino acids, and negative associations of BMI with HDL and HDL cholesterol.[70]

Application and interpretation of findings

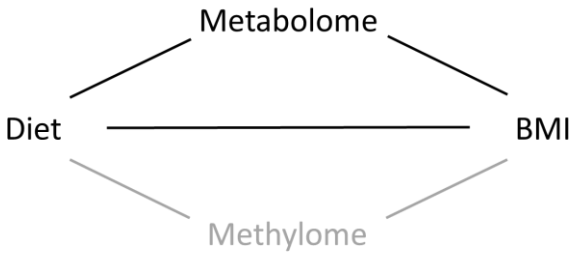
Hivert et al. 2015 discuss the potential for metabolites to be used as markers for complex phenotypes such as dietary intake; this is explored in Chapter 6.[218] They also discuss how metabolomics can be used to help refine phenotypes associated with obesity, for example, insulin resistance and type 2 diabetes risk. Chen et al. compared the metabolic profiles of people with metabolic healthy obesity (MHO) and metabolic abnormal obesity (MAO).[77] MHO is defined as obesity without hyperglycaemia, hypertension or dyslipidaemia, whilst MAO is defined as having at least one metabolic abnormality. Identifying metabolites that differ between people with MHO and MAO may help the discovery of underlying mechanisms that lead to metabolic dysregulation.

Metabolite profiles are strongly associated with BMI. Metabolite profiling can be used to improve identification of groups of people that are at high risk of developing life-threatening and largely preventable diseases such as coronary heart disease. Using metabolomics to spot early warning signs of such diseases could help healthcare professionals to introduce interventions earlier and potentially save lives.

Previous studies in both young adults and older adults have applied Mendelian randomization and found evidence that points to BMI having a causal effect on the metabolome.[70,75] A longitudinal study observed associations between change in BMI and changes in the metabolome.[70] This suggests that lifestyle changes aimed at lowering (or increasing) BMI may also have a positive impact on the metabolite profile.

The above findings in ALSPAC children show that the ability of BMI to influence the metabolome starts in childhood (**Figure 15** and **Figure 16**). Therefore, healthy weight interventions should start in childhood, not only to put in place good lifestyle habits at a young age, but also to establish a healthy metabolite profile.

CHAPTER 6. DIET, METABOLOME AND BMI



6.1. Introduction

Many studies have observed associations between dietary patterns and blood metabolite profiles.[59-65] One application of understanding the relationship between diet and metabolites could be to predict an individual's dietary pattern given their metabolic profile.[64,66,67] This would be beneficial since it is difficult to accurately assess dietary intake in populations.

In order to investigate the relationship between dietary patterns and metabolite profiles some studies looked at habitual dietary patterns. For example, a cross-sectional analysis in adults identified an association between a Western dietary pattern (refined grains, sweet food and processed meat) and increased levels of amino acids, including xleucine (combined leucine and isoleucine) and phenylalanine.[59] Metabolites were quantified using mass spectrometry.

Some studies of habitual dietary patterns tried to summarise an individual's overall dietary pattern using principal components (PCs) or other summary methods,[59-61,63] whereas other studies looked at intakes of individual food groups.[65] A study in women of metabolites (measured using mass spectrometry) and dietary intake patterns (PCs derived from FFQs) identified positive associations between fruit and vegetable intake and several phosphatidylcholines.[63] Some studies looked at metabolites individually or by class,[60,65] whilst other studies used principal components to summarise metabolite profiles.[59,61]

A study conducted in the European Prospective Investigation into Cancer and Nutrition aimed to identify serum metabolites that may relate red meat intake to type 2 diabetes (T2D) risk.[219] They observed associations between red meat intake and 21 metabolites (glycine, 17 phosphatidylcholines, 2 sphingomyelins and ferritin). 13 of these metabolites were also associated with T2D risk, and, of those 13, the direction of effect for 6 of them was consistent with the red meat-metabolite association. Further analysis was performed to investigate whether these 6 metabolites (glycine, 3 phosphatidylcholines, 1 sphingomyelin and ferritin) mediate the relationship between

red meat intake and T2D risk. For each of these 6 metabolites, the association between red meat intake and T2D risk was largely attenuated after adjustment for them, which is consistent with the hypothesis that they mediate the relationship between red meat intake and T2D risk.

Other studies of the relationship between diet and metabolites required participants to follow specific diets for the study.[62,220] A randomized clinical trial compared serum amino acids between infants fed a lower protein formula and infants fed a higher protein formula.[62] They observed that, compared to the lower protein group, infants in the higher protein group had higher serum concentrations of several amino acids including isoleucine, leucine, valine, phenylalanine and tyrosine.

Short-term dietary intervention studies have identified several biomarkers of food intake, however these have mostly been conducted using urine, and the biomarkers identified are often short-term and rapidly excreted.[220]

In summary, several studies have observed associations between dietary intake and metabolite levels, including associations with amino acids and phosphatidylcholines.

Most studies investigating the relationship between habitual diet and a broad range of blood metabolites have used mass spectrometry, not NMR, to measure metabolite levels. NMR-based metabolomics studies of diet have tended to use metabolite levels measured in urine or have focussed on an individual metabolite.[221]

Many strong associations between BMI and the metabolome in the ALSPAC children were observed in the previous chapter. It is likely that some of these BMI-associated metabolites are also associated with dietary behaviour. It is also plausible that some metabolites may mediate the relationship between diet and BMI, or that BMI may mediate the relationship between diet and metabolites. This chapter aims to explore these hypotheses.

6.2. Methods

6.2.1. Diet and the metabolome – cross-sectional analyses

The cross-sectional relationship between dietary behaviour and metabolites was explored in the ALSPAC children. Analyses were performed in the ALSPAC children at age 7 years, but not at age 15 years since diet data are not available at age 15.

Dietary behaviour was measured using PCs generated from FFQs and diet diaries (**2.1.1.1**). Of the six PCs generated from diet data at age 7 years (three PCs from FFQs and three PCs from diet diaries), only two were associated with BMI (**4.3.2.1**). These were the "health aware" PC and the "packed lunch" PC, both generated from the diet diary data.

Metabolite profiles were generated from serum samples and quantified using a NMR platform (**2.1.1.6**). Metabolite data was prepared as described in Chapter 5: metabolites with skewed distributions were normalized, and all metabolites were scaled to standard deviation units.

The main motivation for these analyses is to explore whether the metabolome mediates the effect of dietary intake on BMI, hence the relationship between diet and metabolites was only explored for diet PCs and metabolites known to be associated with BMI in the ALSPAC 7-year-olds.

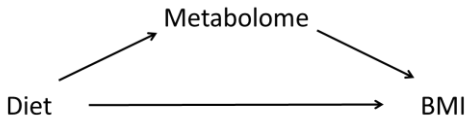
Linear regression was performed to test the associations between each of the diet PCs and metabolites. Models were adjusted for age at serum sampling for metabolite measurement (clinic visit) and sex.

```
lm(metabolite ~ diet + age + sex)
```

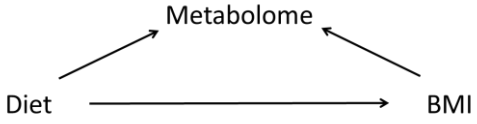
When detecting associations between diet and metabolites, a p-value threshold of $p<0.001$ was used since this threshold was also used in the BMI and metabolome analyses in Chapter 5.

6.2.2. Diet, BMI and the metabolome – analyses

The original question posed in this thesis is whether the metabolome (and the methylome) mediates the effect of diet on BMI. This hypothesis is represented in the following diagram:



However, analyses investigating whether BMI has a causal effect on the metabolome have found evidence suggesting that BMI has a causal effect on several metabolites (**5.4**). Würtz et al. performed MR analyses and found evidence suggesting the BMI has a causal effect on multiple metabolites.[70] The MR analyses in Chapter 5 also suggest that BMI has a causal effect on several metabolites. Therefore, the hypothesis represented in the following diagram is also plausible:

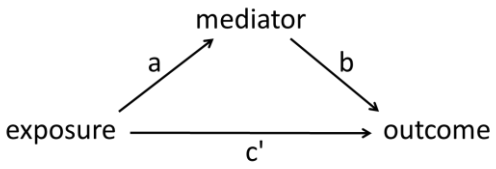


6.2.2.1 Identifying potential mediators

A mediation model with a single mediator can be represented by the diagram in **Figure 17**, where ab is the mediated effect, c' is the direct effect, and the total effect is c = ab + c'.[222] A mediation model is described as "consistent" if the mediated effect (ab) has the same sign as the direct effect (c'), and "inconsistent" if the mediated effect has an opposite sign to the direct effect.[223] For simplicity, this chapter focuses on consistent mediation models.

The total effect is the sum of the mediated effect and the direct effect. Hence, if the total effect has an opposite sign to the mediated effect, then the mediated effect must have an opposite sign to the direct effect, and therefore the mediation model is inconsistent.

Figure 17 – Causal mediation model with a single mediator.
The mediated effect is ab and the direct effect is c'. The total effect is c = ab + c'.



6.2.2.2 Bidirectional MR – BMI and metabolites

The causal effect of BMI on metabolites in ALSPAC 7-year-olds was explored in Chapter 5, but the causal effect of metabolites on BMI was not explored due to a likely lack of power for that analysis. To investigate the hypotheses above (**6.2.2.1**), it is helpful to gain a clearer picture of causality in the relationships between BMI and metabolites. Causality is only investigated here for metabolites which are potential mediators between diet PCs and BMI.

Since a bidirectional MR analysis between BMI and metabolites in ALSPAC is likely to lack power, a two-sample bidirectional MR analysis was performed instead. Two-sample MR analyses were performed in the MR-base web-app (http://www.mrbase.org)[131] using the IVW method.[129,130] BMI GWAS summary results used were from a GWAS conducted in c.320,000 individuals of European descent by Locke et al.[52] Metabolite GWAS summary results used were from a GWAS conducted in c.25,000 individuals from European cohorts by Kettunen et al.[224]

6.2.2.3 Mediation analyses

Mediation analyses were performed to investigate whether metabolites mediate the relationship between diet and BMI, and whether BMI mediates the relationship between diet and metabolites.[17] These analyses were performed using the *mediate* function from the *mediation* package in R (version 3.3.3), which estimates the mediated effect, the direct effect and the total effect of the exposure on the outcome (**2.2.5**).[133] Analyses were adjusted for age and sex.

6.2.2.4 Diet, amino acids and BMI

The effect directions of the pairwise associations between the diet PCs, amino acids and BMI fit with an inconsistent mediation model rather than a consistent mediation model (described in **6.2.2.1**). However, it is interesting that three of the amino acids showed strong associations with both diet PCs. To explore how BMI relates to these associations, the children were split into groups according to their BMI quartile, and simple linear regression models (amino acid ~ diet PC) were fitted and plotted.

6.3. Results

6.3.1. Diet and metabolome – cross-sectional results

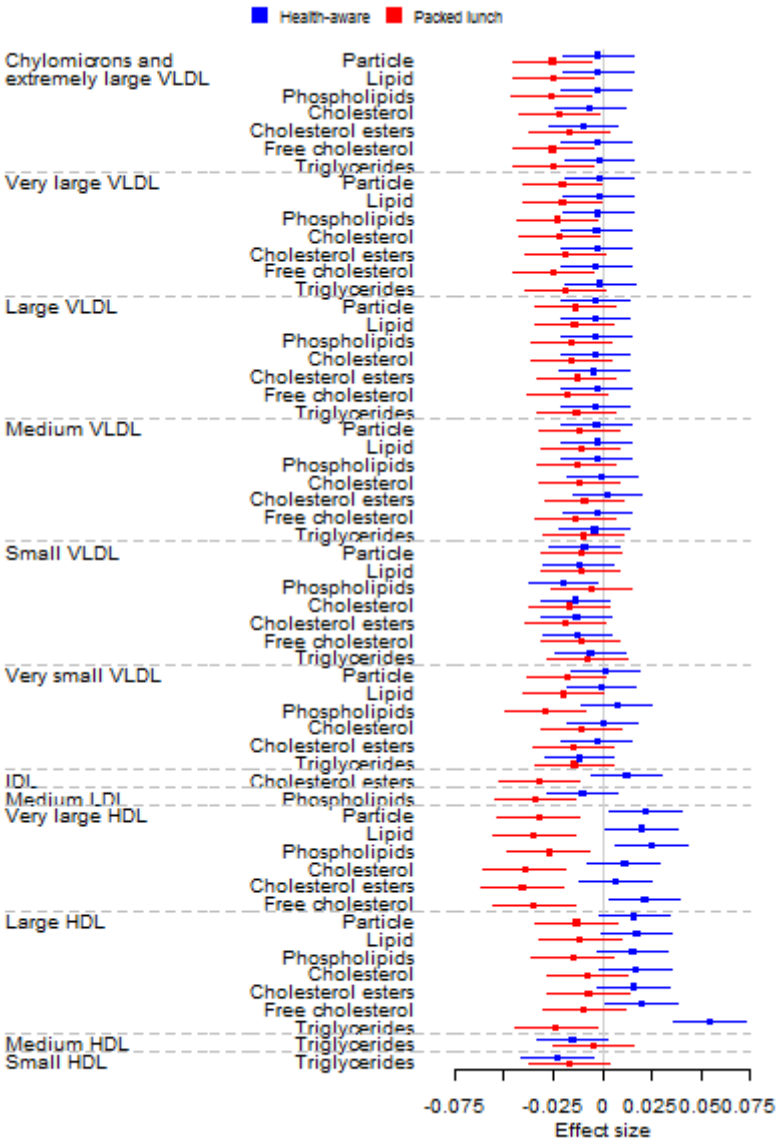
Cross-sectional analyses were performed to investigate the relationship between the diet PCs and metabolites at age 7 years. The results of these analyses are presented as a forest plot in **Figure 18** for the BMI-associated metabolites, and as a table in the appendix for all metabolites.

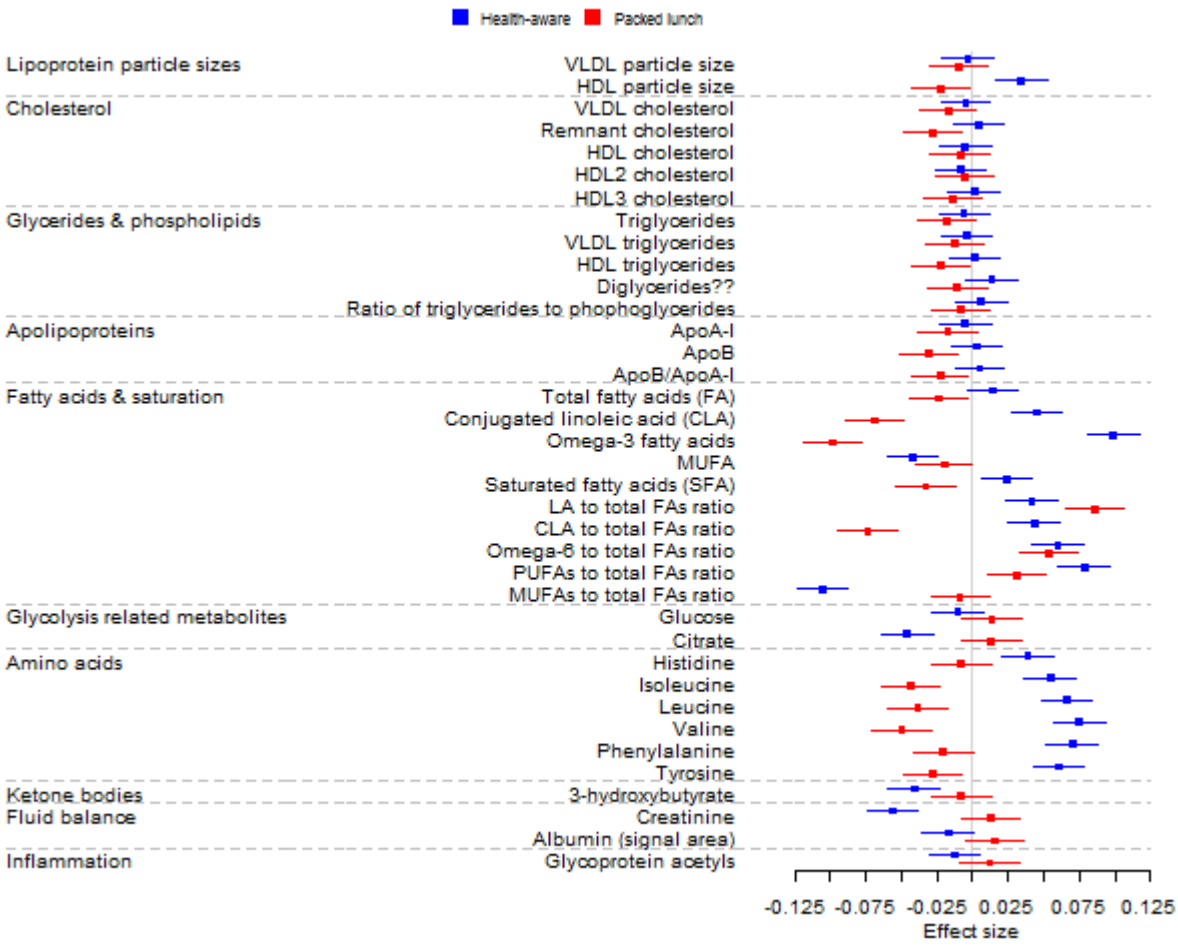
Using a p-value threshold of 0.001 (since this threshold was used in analyses of BMI and the metabolome in Chapter 4), the health aware PC was positively associated with some large HDL triglycerides, HDL particle diameter, CLA, omega-3, histidine, isoleucine, leucine, valine, phenylalanine and tyrosine. It was negatively associated with MUFA, citrate, 3-hydroxybutyrate and creatinine.

The packed lunch PC was negatively associated with medium LDL phospholipids, very large HDL cholesterol and cholesterol esters, CLA, omega-3, isoleucine, leucine and valine.

Figure 18 – Forest plots of cross-sectional relationships between metabolites and diet PCs in the ALSPAC children at age 7 years.

Effect estimates are the 1-SD increase in metabolite concentration per unit increase in diet PC. The lines though the points are the 95% CIs. The metabolites have been divided into two plots. The first plot (below) shows the results for the fourteen lipoprotein subclasses and their lipid measures, and the second plot (following page) shows results for the remaining metabolite measures.




Health-aware
Packed lunch
Lipoprotein particle sizes
VLDL particle size
HDL particle size
Cholesterol
VLDL cholesterol
Remnant cholesterol
HDL cholesterol
HDL2 cholesterol
HDL3 cholesterol
Glycerides & phospholipids
Triglycerides
VLDL triglycerides
HDL triglycerides
Diglycerides??
Ratio of triglycerides to phophoglycerides
Apolipoproteins
ApoA-I
ApoB
ApoB/ApoA-I
Fatty acids & saturation
Total fatty acids (FA)
Conjugated linoleic acid (CLA)
Omega-3 fatty acids
MUFA
Saturated fatty acids (SFA)
LA to total FAs ratio
CLA to total FAs ratio
Omega-6 to total FAs ratio
PUFAs to total FAs ratio
MUFAs to total FAs ratio
Glycolysis related metabolites
Glucose
Citrate
Amino acids
Histidine
Isoleucine
Leucine
Valine
Phenylalanine
Tyrosine
Ketone bodies
3-hydroxybutyrate
Fluid balance
Creatinine
Albumin (signal area)
Inflammation
Glycoprotein acetyls
-0.125 -0.075 -0.025 0.025 0.075 0.125
Effect size

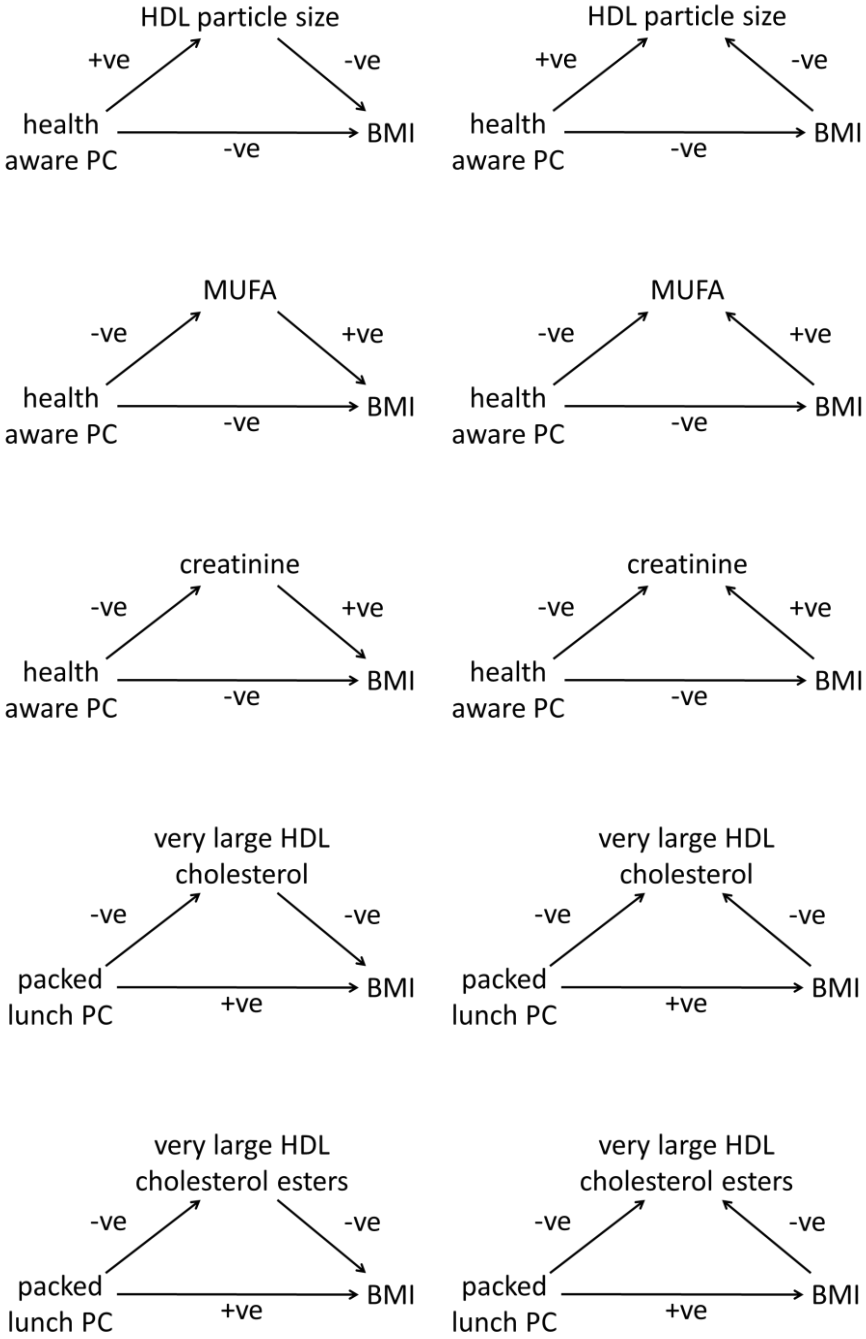
6.3.2. Diet, BMI and the metabolome – results

6.3.2.1 Potential mediators

Observed cross-sectional associations for five of the metabolites meet the conditions for consistent mediation models described in **6.2.2.1** and **Figure 17** above. For the health aware PC these metabolites are HDL particle size, MUFA and creatinine. For the packed lunch PC these are very large HDL cholesterol and cholesterol esters. The diagrams in **Figure 19** represent the mediation hypothesis to be explored for these five metabolites. The diagrams on the left-hand side of this figure represent the hypothesis that the metabolites mediate the effect of diet on BMI, and the diagrams on the right-hand side represent the hypothesis that BMI mediates the effect of diet on the metabolites.

For the mediation analyses conducted in this chapter, it is assumed that the diet PCs have causal effects on both BMI and metabolites. Causal effects between BMI and the metabolites are explored in **6.3.2.2**.

Figure 19 – Diagrams representing the mediation hypothesis to be explored.



6.3.2.2 Two-sample bidirectional MR of metabolites and BMI in MR-base

Two-sample bidirectional MR analyses were performed to investigate causality between BMI and the five metabolites identified as potential mediators in **6.3.2.1**. Results of these analyses are reported in **Table 16**. The aim of these analyses was to assess whether it seems more likely that the metabolites mediate the effect of diet on BMI or that BMI mediates the effect of diet on the metabolites.

MR results suggest a causal effect of BMI on HDL particle size – a unit increase in BMI is associated with a 0.314 SD decrease in mean diameter of HDL particles ($p=6.32 \times 10^{-6}$). Weak evidence (given multiple testing) was observed for a causal effect of BMI on very large HDL (XLHDL) cholesterol and XLHDL cholesterol esters – a unit increase in BMI is associated with a 0.159 SD decrease in XLHDL cholesterol (p=0.011) and a 0.155 SD decrease in XLHDL cholesterol esters (p=0.015). These effects estimates are directionally consistent with the observational estimates (**Figure 15**). MR did not detect a causal effect of the metabolites on BMI.

This contrasts with the BMI to metabolite MR analysis in the ALSPAC children (**Figure 20**), which suggested that BMI has a causal effect on creatinine but did not detect a causal effect of BMI on very large HDL cholesterol and cholesterol esters. The very large HDL cholesterol and cholesterol esters confidence intervals observed here, however, do overlap with the confidence intervals observed in the ALSPAC children.

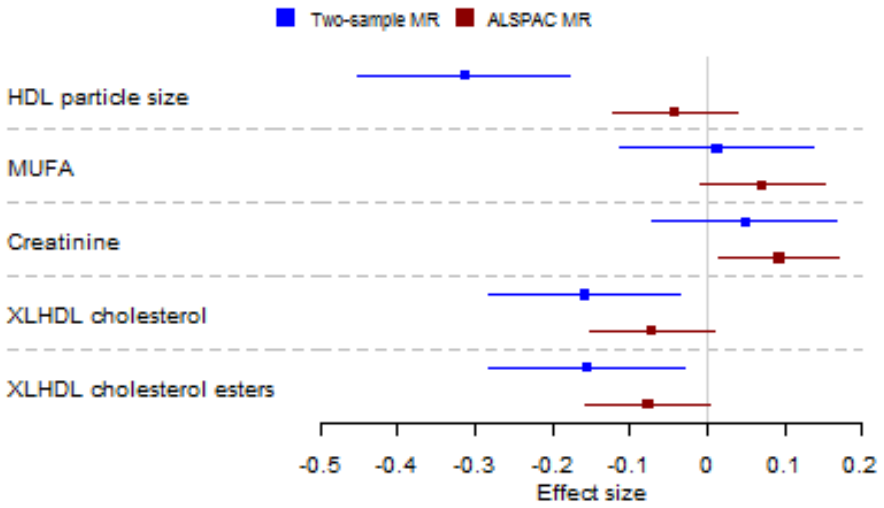
Table 16 – Results from two-sample bidirectional MR analyses.

Effect estimates for the metabolite to BMI MR analyses are the increase in BMI (kg/m^2) per 1-SD increase in metabolite concentration. Effect estimates for the BMI to metabolite MR analyses are the increase in metabolite concentration (SD) per unit increase in BMI (kg/m^2). XLHDL, very large HDL.

Exposure	No. SNPs	Outcome	Beta	95% CI	p-value
HDL particle size	10	BMI	0.002	-0.021, 0.025	0.869
MUFA	5	BMI	-0.031	-0.103, 0.042	0.406
Creatinine	5	BMI	0.022	-0.037, 0.082	0.464
XLHDL cholesterol	8	BMI	-0.005	-0.036, 0.026	0.752
XLHDL cholesterol esters	7	BMI	-0.014	-0.042, 0.014	0.330
BMI	68	HDL particle size	-0.314	-0.451, -0.178	6.32×10^{-6}
BMI	68	MUFA	0.012	-0.113, 0.137	0.848
BMI	68	Creatinine	0.048	-0.070, 0.167	0.426
BMI	68	XLHDL cholesterol	-0.159	-0.281, -0.036	0.011
BMI	68	XLHDL cholesterol esters	-0.155	-0.281, -0.030	0.015

Figure 20 – Forest plot comparing BMI → metabolite MR estimates from the two-sample MR analysis with those from the ALSPAC MR analysis.

ALSPAC results are from the single sample MR analysis conducted in the ALSPAC children at age 7. Effect estimates are the increase in metabolite concentration (SD) per unit increase in BMI (kg/m^2). XLHDL, very large HDL.



6.3.2.3 Diet, BMI and metabolites – mediation analysis results

Mediation analyses were conducted to investigated whether the metabolites mediate the relationship between the diet PCs and BMI, and whether BMI mediates the relationship between the diet PCs and the metabolites. The mediated effect (ab in **Figure 17**), the direct effect (c' in **Figure 17**) and the total effect (the sum of the mediated effect and the direct effect) were estimated.

Results from the mediation analyses investigating whether the metabolites mediate the effect of the diet PCs on BMI are presented in **Figure 21**. These results suggest that HDL particle size, MUFA and creatinine mediate the relationship between the health aware PC and BMI and that very large HDL cholesterol and cholesterol esters mediate the relationship between the packed lunch PC and BMI.

Results from the mediation analyses investigating whether BMI mediates the effect of the diet PCs on metabolites are presented in **Figure 22**. These results suggest that BMI mediates the relationship of the health aware PCs with very large HDL cholesterol and

cholesterol esters. Weaker evidence was observed for BMI as a mediator between the health aware PC and HDL particle size, MUFA and creatinine.

In most of these analyses the magnitude of mediated effect estimate is much less than the magnitude of the direct effect estimate. The exception to this is the analysis of creatine as a mediator between the health aware PC and BMI, in which the mediated effect estimate is c. twice the size of the direct effect estimate.

Figure 21 – Forest plot of results from mediation analyses exploring whether the metabolites mediate the effect of the diet PCs on BMI.

Each row represents a mediation analysis and shows the effect estimates for the mediated, direct and total effects calculated in that mediation analysis. The effect estimate units are the increase in BMI (kg/m^2) per unit increase in diet PC score. Rows labels are the particular diet PC and metabolite that were studied in that mediation analysis. P-values for the mediated effects estimates are all <0.01. HDLD, HDL particle diameter; MUFA, monounsaturated fatty acids; Crea, Creatine; XLHDLC, very large HDL cholesterol; XLHDLC, very large HDL cholesterol esters.

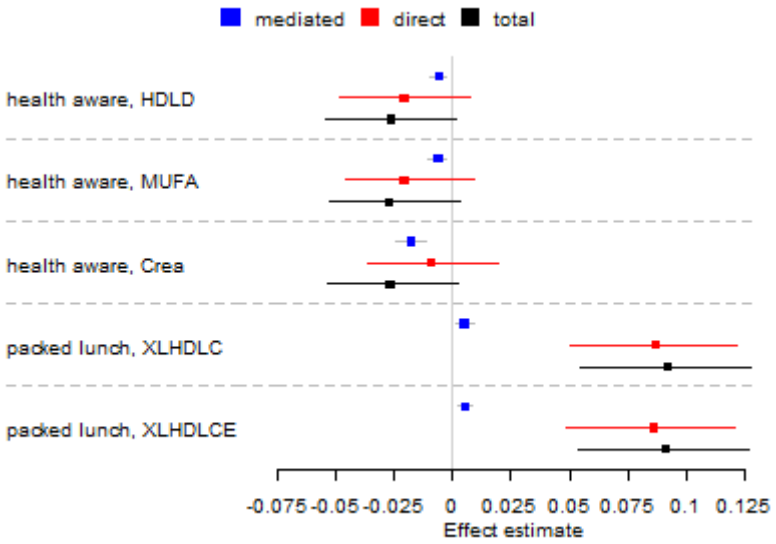
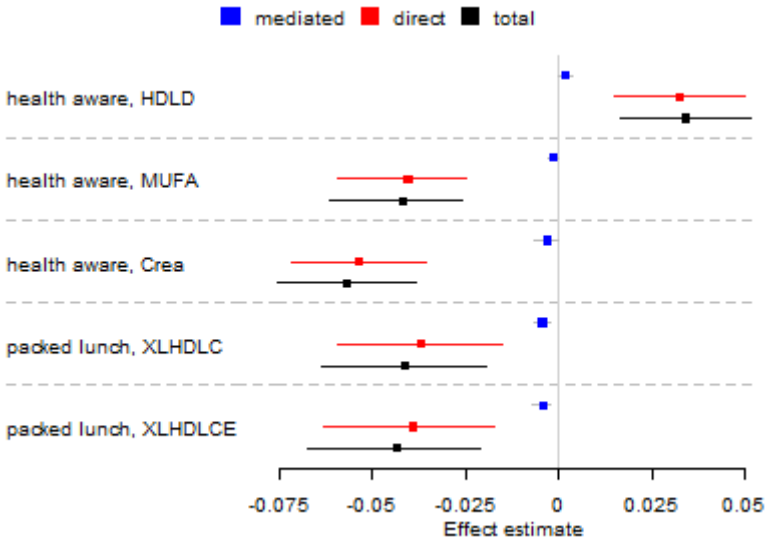


Figure 22 – Forest plot of results from mediation analyses exploring whether BMI mediates the effect of the diet PCs on the metabolites.

Each row represents a mediation analysis and shows the effect estimates for the mediated, direct and total effects calculated in that mediation analysis. The effect estimate units are the increase in diet PC score per unit increase in BMI (kg/m^2). Rows labels are the particular diet PC and metabolite that were studied in that mediation analysis. P-values for the mediated effects estimates are ~0.05-0.06 in the analyses using the health aware PC and <0.01 in the analyses using the packed lunch PC.

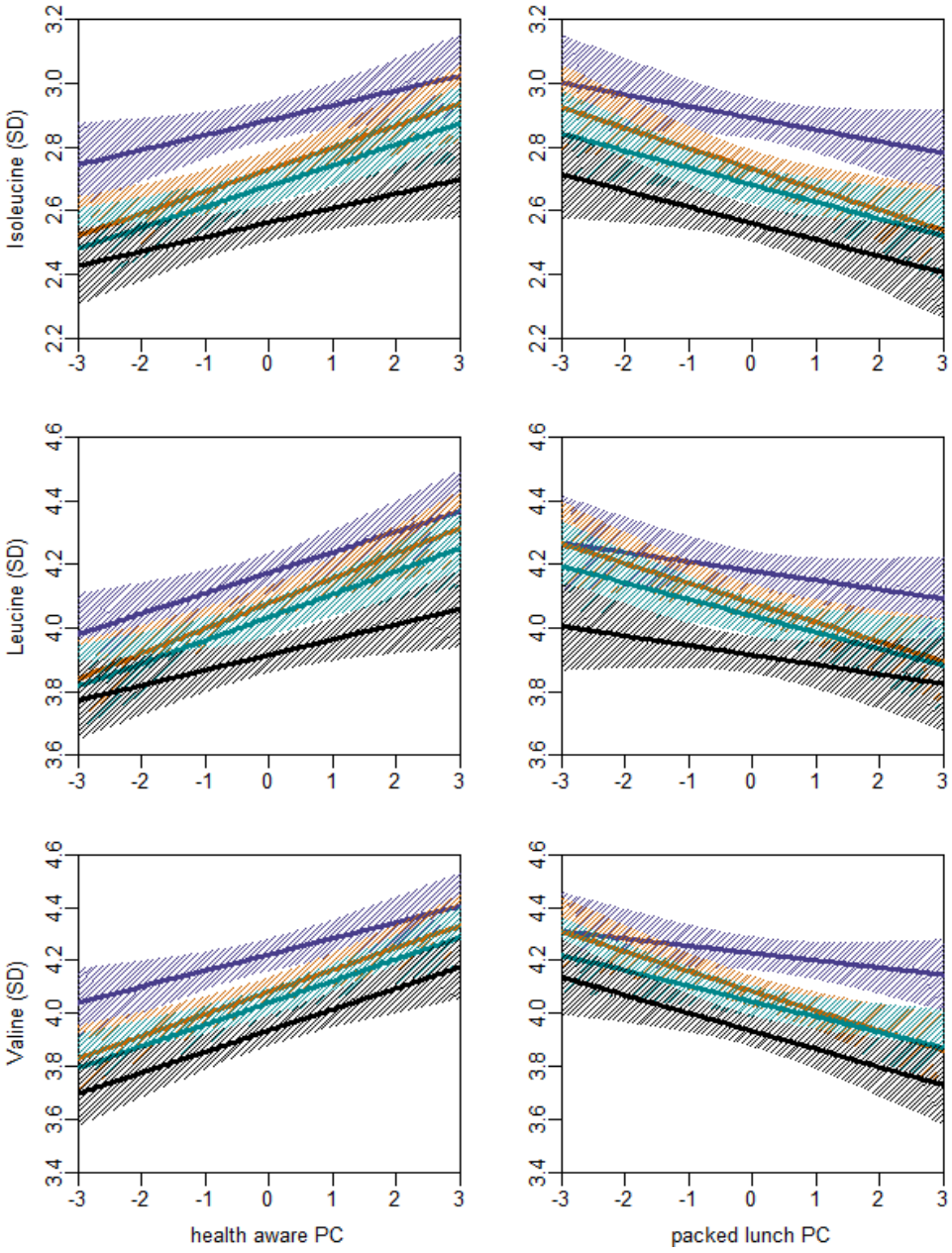


6.3.2.4 Diet and metabolite relationship by BMI quartile

The role of BMI in the relationship between three amino acids (isoleucine, leucine and valine) and diet PCs was explored by splitting the children into groups according to their BMI quartile and fitting linear regression models (amino acid ~ diet PC) for each amino acid and diet PC pair. The results of these analyses are plotted in **Figure 23**. In these analyses, as amino acid levels increase, the health aware PC increases and the packed lunch PC decreases. Also, as amino acid levels increase, BMI increases.

Figure 23 – Diet and metabolite lines of best fit by BMI quartile.

Black line for quartile with BMI between 13 and 14.9, turquoise line for BMI between 14.9 and 15.8, orange line for BMI between 15.8 and 17, blue line for BMI between 17 and 21.5 kg/m^2. Shaded areas are the 95% confidence intervals.



6.4. Discussion

This chapter explored mediation in the relationships between dietary behaviour, metabolites and BMI.

Metabolites and diet

Analyses in the previous chapter identified BMI-associated metabolites in the ALSPAC children. Analyses conducted in this chapter found that several of the BMI-associated metabolites are also associated with dietary behaviour.

As described in Chapter 2 (**2.1.1.1**), the health aware and packed lunch PCs were derived from diet diary data from the ALSPAC children at age 7 years. The health aware PC has positive loadings for intake of cheese, high fibre bread, pasta, salad, fresh fruit and fruit juice, and negative loadings for intake of processed meat, chips and diet fizzy drinks; the packed lunch PC has positive loadings for intake of low fibre bread, margarine, ham, bacon, crisps and diet squash (Northstone et al., unpublished).

Conjugated linoleic acid (CLA) and omega-3 fatty acids were positively associated with the health aware PC and negatively associated with the packed lunch PC. The main sources of omega-3 fatty acids are fish and seafood;[225] dietary sources of CLA include dairy products and beef.[226] The health aware PC is associated with a higher intake of cheese (a source of CLA) (Northstone et al., unpublished).

The health aware PC displayed positive associations with several amino acids (histidine, isoleucine, leucine, valine, phenylalanine and tyrosine). The packed lunch PC was negatively associated with branched-chain amino acids (isoleucine, leucine and valine). Histidine, isoleucine, leucine, valine and phenylalanine are essential amino acids, and hence must come from diet since they cannot be synthesized *de novo* in humans. Previous studies have observed positive associations of amino acids with a Western dietary pattern (high in refined grains, sweet food, processed meat) and a dietary pattern high in potatoes, dairy products, vegetables and cornflakes.[59,60] In ALSPAC, the health aware PC is associated with higher intakes of cheese and lower intakes of

processed meat and the packed lunch PC is associated with higher intakes of ham and bacon (Northstone et al., unpublished).

The health aware PC was also positively associated with some large HDL triglycerides and HDL particle diameter and negatively associated with MUFA, citrate, 3-hydroxybutyrate and creatinine. The packed lunch PC was negatively associated with very large HDL cholesterol and cholesterol esters.

Mendelian randomization

Two-sample MR analyses exploring the effect of BMI on metabolites observed evidence of a negative causal effect of BMI on HDL particle size and very large HDL cholesterol and cholesterol esters but did not detect a causal effect of BMI on MUFA or creatinine (**Table 16**). This is mostly consistent with Würtz et al. who observed a negative causal effect of BMI on HDL particle size, a weak negative causal effect of BMI on MUFA, and no causal effect of BMI on creatinine.[70] In contrast, the single sample MR analysis in Chapter 5 (**5.3.2**) observed a positive causal effect of BMI on creatinine, but did not detect a causal effect of BMI on HDL particle size, MUFA or very large HDL cholesterol or cholesterol esters. This inconsistency in identifying a causal effect of BMI on creatinine may be due to differing observational associations between creatinine and BMI in different age groups. Strong positive associations were observed between creatinine and BMI in the ALSPAC children and teenagers, but no association was found in the Würtz young adults (**Figure 13**).

Two-sample MR analyses exploring the effects of five metabolites (HDL particle size, MUFA, creatinine and very large HDL cholesterol and cholesterol esters) on BMI did not detect any causal effects (**Table 16**). This lack of evidence may be due to an absence of causality, or it may be due to the genetic instruments lacking the strength to identify these effects.

Mediation

Based on observed pairwise associations between the diet PCs, metabolites and BMI, five metabolites were identified as potential mediators in the relationship between diet

and BMI (or that the relationship between diet and those metabolites may be mediated by BMI). These metabolites were HDL particle size, MUFA and creatinine for the health aware PC, and very large HDL cholesterol and cholesterol esters for the packed lunch PC.

Mediation analyses were performed to investigate whether metabolites mediate the relationship between diet and BMI, and whether BMI mediates the relationship between diet and metabolites. Results for the packed lunch PC suggest that very large HDL cholesterol and cholesterol esters mediate the relationship between the packed lunch PC and BMI. The results also suggest that BMI mediates the relationship between the packed lunch PC and very large HDL cholesterol and cholesterol esters. In both cases the mediated effect estimate is much smaller than the direct effect.

Weak evidence was observed for BMI as mediator between the health aware PC and HDL particle size, MUFA and creatinine ($p \approx 0.05$-0.06). Results suggest that HDL particle size, MUFA and creatinine mediate the relationship of the health aware PC with BMI.

Amino acids

Isoleucine, leucine and valine were positively associated with the health aware PC and negatively associated with the packed lunch PC, which suggests that they are strongly linked to a range of dietary behaviour. However, the pairwise effect directions between these metabolites, the diet PCs and BMI did not fit the hypothesis of a consistent mediation model (**6.2.2.1**).

Suppose a child scores highly for the health aware PC. Then, according to the findings in this chapter, one would expect them to have higher levels of valine and hence a higher BMI. However, according to findings in Chapter 4, one would expect them to have a lower BMI. One explanation for this could be that diet influences BMI through two pathways. A child may have a high health aware score because a large proportion of their diet consists of healthy foods, or a child may have a high health aware score because they consume large quantities of food, including healthy food and less healthy food. If a child has a high health aware score because much of their diet consists of healthy food, then one would expect them to have a high valine intake but a low energy

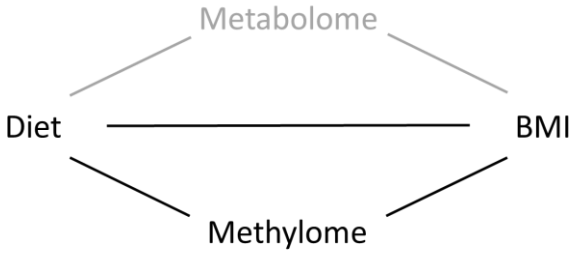
intake, and hence a lower BMI. If a child has a high health aware score because they have a high overall food intake, then one would expect them to have a high valine intake and a high energy intake, and hence a higher BMI.

Conclusions

Bringing together findings from the MR analyses and mediation analyses, the results suggest that BMI mediates the effect of the packed lunch PC on very large HDL cholesterol and cholesterol esters (assuming the packed lunch PC has a causal effect on BMI and these metabolites). Results also suggest that BMI mediates the effect of the health aware PC on HDL particle size and creatinine (assuming the health aware PC has a causal effect on BMI and HDL particle size). Causal inference in these mediation analyses is limited by the lack of genetic instruments for the diet PCs, hence causal effects of the diet PCs on BMI are assumed but not tested.

Strong links have been observed between dietary patterns, the metabolome and BMI. The metabolome plays a key role when trying to understand the relationship between dietary behaviour and BMI. Some of these relationships may change across the lifecourse, for example the association between creatinine and BMI.

CHAPTER 7. BMI, METHYLATION AND DIET



7.1. Introduction

Variation in DNA methylation has been linked to adiposity and (less strongly) to dietary behaviour. This chapter explores whether previously identified association between methylation and BMI in adults also hold in the ALSPAC children and adolescents. This chapter also seeks to identify novel associations for methylation with diet and BMI in the ALSPAC children and adolescents.

7.1.1. BMI and methylation

The first major EWAS to report robust associations between CpG sites and BMI was a study by Dick et al. in 2014.[91] They identified an association between increased methylation at three CpGs (cg22891070, cg27146050 and cg16672562) in *HIF3A* and increased BMI in their study of adults of European origin. This widely cited paper made an initial foray in to deciphering the direction of the causal relationship between DNA methylation variation at the *HIF3A* locus but fell short of formally applying Mendelian randomization to strengthen causal inference. We subsequently pursued this line of enquiry to show that the direction of the causal pathway was most likely to go from BMI to DNA methylation. The study in ALSPAC of *HIF3A* methylation and BMI in this chapter describes work published in Diabetes in which I was joint first author (Richmond RC, Sharp GC, Ward ME, et al. DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework. *Diabetes* 2016; **65**(5): 1231-44.). Analyses were performed by myself, Dr Gemma Sharp and Dr Rebecca Richmond. The full manuscript can be found in **Appendix B** of this thesis.

More recently, Wahl et al. performed an EWAS of BMI in 10,261 adults and identified 187 CpGs associated with BMI at an epigenome-wide level (defined as $p < 1 \times 10^{-7}$ here).[94] In their discovery analyses they studied 5,387 adults of European (n = 2,707) and Indian Asian (n = 2,680) ancestry and identified 287 BMI-associated CpGs ($p < 1 \times 10^{-7}$) across 207 genetic loci. They adjusted their analyses for age, sex, smoking status, physical activity index, alcohol consumption. They took the top CpG at each locus (the CpG with the lowest p-value) forward for replication in 4,874 European and Indian Asian

adults. They found that 187 of the 207 CpGs replicated, defining replication criteria as a directionally consistent effect estimate and $p < 0.05$ in the replication samples and $p < 1 \times 10^{-7}$ in the meta-analysis of the discovery and replication cohorts. The genetic loci identified by these 187 CpGs include genes involved in lipid and lipoprotein metabolism.

Mendelson et al. performed a BMI EWAS in 7,798 adults and identified 83 BMI-associated CpGs.[95] In their discovery analyses they studied 3,743 adults from three American and Scottish cohorts; and in their replication analyses they studied 4,055 adults of African and European ancestry. They identified 135 BMI-associated CpGs in their discovery analyses (Bonferroni-corrected threshold of $p < 1.2 \times 10^{-7}$), of which 83 replicated ($p < 0.05/135$ in the meta-analysis of the replication cohorts). They adjusted their analyses for age and sex. Mendelson et al. studied whole blood gene expression and identified associations between expression of genes in lipid metabolism pathways and BMI-associated CpGs.

Both Wahl et al. and Mendelson et al. measured DNA methylation from blood samples using the Illumina Infinium HumanMethylation450 BeadChip. Out of the 187 and 83 BMI-associated CpGs identified in their respective EWAS, 38 CpGs were common to both analyses.

Wahl et al. used MR to investigate causality between BMI and their 187 BMI-associated CpGs. One CpG (cg2666590 at the *NFATC2IP* gene locus) appeared to have a causal effect on BMI, whilst BMI appeared to have a causal effect on three CpGs (cg00138407, cg06500161 and cg09613192 at the *KLHL18*, *ABCG1* and *FTH1P20* gene loci respectively). Mendelson et al. also conducted MR to investigate causality between BMI and methylation. Out of their 83 BMI-associated CpGs, their results suggest that one CpG (cg11024682 at the *SREBF1* gene locus) has a causal effect on BMI and that BMI has a causal effect on 16 CpGs (including cg06500161 that was also identified in the BMI → methylation MR analysis by Wahl et al.), though they used less stringent p-value thresholds than Wahl et al. to infer causality. Both studies agreed that the prevailing weight of evidence supported the hypothesis that changes in methylation mostly seem

to be a consequence of changes in BMI, rather than a cause. This is consistent with our own previous analyses investigating causality between *HIF3A* methylation and BMI.[109]

7.1.2. Diet and methylation

Few EWAS have been able to identify robust associations between dietary behaviour and DNA methylation.

An EWAS of tea and coffee consumption in 3,096 adults from four European cohorts identified two CpGs associated with tea consumption in women, however these associations did not hold in the men-only or sex-combined analyses.[85] No epigenome-wide significant associations were identified in the men-only or sex-combined analyses.

An EWAS of a Mediterranean-style dietary pattern in 3,563 Framingham Heart Study participants identified an association between a single CpG (cg05575921 in the *AHRR* gene) and the dietary pattern.[86] However, this CpG site has been very widely and robustly associated with exposure to tobacco smoke[227,228] and therefore the association is highly likely to be explained by residual confounding. They also studied the relationship between cg05575921 and components of the dietary pattern and observed an association with fruit and whole grain intake.

An earlier study of global DNA methylation observed evidence of an association for a high fruit and vegetable intake with a lower prevalence of global hypomethylation.[229]

7.1.3. Motivation for these analyses

Previous epigenome-wide studies of the relationship between BMI and methylation have identified several BMI-associated CpGs in adulthood but have not investigated whether these associations were also present in childhood and adolescence. These analyses therefore aim to identify cross-sectional associations between BMI and methylation in childhood and adolescence. These analyses also aim to investigate whether previously identified associations between BMI and methylation in adults can be detected earlier in life in children and adolescents.

Analyses will also be performed to investigate the relationship between BMI-associated CpGs and diet, since dietary behaviour may play a role in the relationship between methylation and BMI.

7.2. Methods

7.2.1. *HIF3A* analyses

The relationship between adiposity and methylation at the *HIF3A* loci (cg22891070, cg27146050 and cg16672562) identified by Dick et al. in adults was followed up in ALSPAC in childhood and adolescence.[91] BMI was log-transformed for these analyses so that results could be more easily compared with those of Dick et al.

7.2.1.1 Cross-sectional analyses

Linear regression models were fitted to test the cross-sectional relationships between BMI and methylation at each of the *HIF3A* CpGs in childhood (age 7 years) and adolescence (age 15-17 years). Models were adjusted for age, sex and bisulphite conversion batch in the childhood analyses, and additionally for smoking status in the adolescence analyses:

```
lm(log(BMI) ~ methylation + age + sex + batch)          (age 7)

lm(log(BMI) ~ methylation + age + sex + batch + smoking)  (age 15-17)
```

7.2.1.2 Longitudinal analyses

Linear regression models were fitted to test the association between BMI at age 7 and each of the *HIF3A* CpGs at age 15-17, and between BMI at age 15-17 and each of the *HIF3A* CpGs at age 7. Models were adjusted for age, sex and bisulphite conversion batch:

```
lm(log(15-17y BMI) ~ 7y methylation + age + sex + batch)

lm(15-17y methylation ~ log(7y BMI) + age + sex + batch)
```

Models were also fitted with additional adjustment for baseline methylation/BMI:

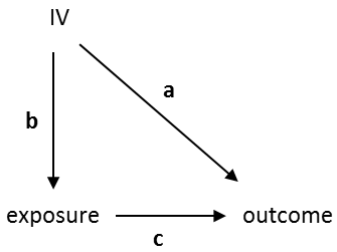
```
lm(log(15-17y BMI) ~ 7y methylation + age + sex + batch + 7y BMI)

lm(15-17y methylation ~ log(7y BMI) + age + sex + batch + 7y methylation)
```

7.2.1.3 Bidirectional MR analyses

Bidirectional MR analyses were performed to explore the causal relationship between BMI and *HIF3A* methylation. For the MR analysis of methylation → BMI, methylation was instrumented using previously identified *cis*-SNPs (rs8102595 and rs3826795), combined in a weighted allele score.[91] For the MR analysis of BMI → methylation, BMI was instrumented using the GIANT BMI allele score.[52]

MR was implemented using the "triangulation" instrumental variable (IV) approach to determine the direction and magnitude of causal effects in the observed associations between BMI and *HIF3A* methylation. The triangulation approach estimates the effect of the IV-outcome association based on the effect estimates for the IV-exposure and exposure-outcome associations.[76] This is compared with the observed effect estimates for the IV-outcome association and a test for difference between the observed and expected estimates is performed (**Figure 24**). The expected effect estimate is calculated by multiplying together the IV-exposure effect estimate and the exposure-outcome association. The standard error for the effect estimate is calculated using a second-order Taylor series expansion of the product of two means.[230]

Figure 24 – The triangulation approach for MR.
The observed effect estimate (**a**) is compared with the expected effect estimate (**b** × **c**).



7.2.2. BMI EWAS in ALSPAC in childhood and adolescence

EWAS analyses were performed in R (version 3.4.1) using the *meffil* R package (**2.2.3**).[123] EWAS regression models were fitted using Independent Surrogate Variable Analysis (ISVA), which models confounding factors as statistically independent surrogate variables.[124] The R script used to conduct the EWAS using *meffil* was compiled by Dr Gemma Sharp. Methylation was calculated as β-values (**2.1.1.5**).[109]

EWAS analyses were performed to investigate the cross-sectional relationship between BMI and methylation at ages 7 and 15-17 years. In these analyses methylation was modelled as the outcome variable, and BMI as the exposure variable. Analyses were adjusted for age, sex, maternal education and Houseman-estimated cell counts (B-cells, CD4+ T-cells, CD8+ T-cells, granulocytes, monocytes and NK).[125] The age 15-17 years analysis was also adjusted for smoking behaviour.

7.2.3. Look-up of previously reported adult BMI CpGs in ALSPAC offspring

EWAS by Wahl et al. and Mendelson et al. have identified 232 CpGs associated with BMI in adults.[94,95] However, they did not investigate whether these associations are also present in children and adolescents. Wahl et al. replicated their findings in the ALSPAC mothers, but not the children.

Results for the 232 adult BMI-associated CpGs were looked up in the results from the EWAS of BMI in childhood and adolescence in ALSPAC.

The relationship between the 232 CpGs and FMI (fat mass index) was also investigated in childhood and adolescence in ALSPAC. Since *meffil* uses data from all available CpGs concurrently to perform functional normalization, the simplest way to calculate the effect estimates for the associations between the 232 CpGs and FMI was to perform an EWAS of FMI and obtain the 232 FMI-CpG effect estimates from the EWAS results. These analyses were adjusted for the same covariates as the BMI EWAS, i.e. for age, sex,

maternal education and Houseman-estimated cell counts in the childhood EWAS, and additionally for smoking behaviour in the adolescent EWAS

7.2.4. Bidirectional MR analyses

Bidirectional MR analyses were performed to investigate causality between BMI and the CpGs that were most strongly associated with BMI in the above EWAS in ALSPAC.

7.2.4.1 BMI to methylation

The causal effect of BMI on BMI-associated CpGs was estimated by modelling the relationship between each of these CpGs and the GIANT BMI allele score.[52]

The effect estimates of the BMI allele score on the BMI-associated CpGs were calculated by performing an EWAS of the BMI allele score at ages 7 and 15-17 years and obtaining from them results for the BMI-associated CpGs. These EWAS were adjusted for the same covariates as in the BMI EWAS above, and also for the first 10 genetic PCs.

7.2.4.2 Methylation to BMI

Two-sample MR analyses were conducted to assess the causal effect of methylation on BMI. Analyses were performed in R (version 3.3.3) using the *mr_singlesnp* function from the *TwoSampleMR* package.[131]

Analyses were performed for each BMI-associated CpG separately. CpGs were instrumented by methylation quantitative trait loci (mQTLs) identified by Gaunt et al.[231]

7.2.5. BMI-associated CpGs and diet PCs

Previous analyses in chapter 4 (**4.3.2.1**) identified associations between BMI and "packed lunch" and "health aware" diet PCs in the ALSPAC children at age 7 years. It was postulated that these BMI-associated dietary behaviours may also be associated with variation in DNA methylation at sites shown to be linked to BMI.

Analyses were conducted to investigate the cross-sectional relationship between BMI-associated CpGs and the "packed lunch" and "health aware" diet PCs in the 7-year-olds. The effect estimates of the diet PCs on the BMI-associated CpGs were calculated by performing separate EWAS of the "packed lunch" PC and the "health aware" PC and obtaining from them results for the BMI-associated CpGs. These EWAS were adjusted age, sex, maternal education and Houseman-estimated cell counts.

7.3. Results

7.3.1. *HIF3A* results

7.3.1.1 Cross-sectional results

A 0.1 unit increase in methylation ß-value at cg27146050 was associated with a 4.66% increase in BMI in adolescence. Methylation at cg27146050 was not associated with BMI in childhood, and methylation at cg22891070 and cg16672562 were not associated with BMI in childhood or adolescence.

Table 17 – Cross-sectional results for BMI and *HIF3A* methylation.
Childhood analyses are adjusted for age, sex and batch. Adolescent analyses are adjusted for age, sex, smoking and batch. Effect sizes are the percentage change in BMI for every 0.1 unit increase in methylation β-value.

CpG	Childhood (N=970)			Adolescence (N=845)		
	% change in BMI	95% CI	p-value	% change in BMI	95% CI	p-value
cg22891070	0.44	-0.35, 1.23	0.27	0.66	-0.31, 1.63	0.19
cg27146050	0.62	-1.69, 2.93	0.60	4.66	1.04, 8.29	0.01
cg16672562	0.31	-0.32, 0.93	0.34	0.40	-0.41, 1.20	0.34

7.3.1.2 Longitudinal results

Longitudinal analyses identified a positive association between childhood BMI and cg27146050 methylation in adolescence – a 10% increase in BMI was associated with a 0.003 increase in methylation (**Table 18**). However, no associations were identified between childhood methylation and adolescent BMI (**Table 19**).

Table 18 – Childhood BMI to adolescent methylation.

Coefficients are change in methylation per 10% increase in BMI.

CpG	Model adjusted for age, sex and batch			Model adjusted for age, sex, batch and childhood methylation		
	Change in methylation	95% CI	p-value	Change in methylation	95% CI	p-value
cg22891070	0.005	-0.002, 0.011	0.17	0.001	-0.004, 0.005	0.78
cg27146050	0.003	0.001, 0.005	0.001	0.003	0.001, 0.004	0.001
cg16672562	0.005	-0.003, 0.013	0.21	0.002	-0.004, 0.008	0.60

Table 19 – Childhood methylation to adolescent BMI.

Coefficients have been converted into percentage change in BMI for every 0.1 unit increase in methylation β-value.

CpG	Model adjusted for age, sex and batch			Model adjusted for age, sex, batch and childhood BMI		
	% change in BMI	95% CI	p-value	% change in BMI	95% CI	p-value
cg22891070	0.68	-0.40, 1.76	0.22	0.14	-0.64, 0.91	0.73
cg27146050	2.30	-0.83, 5.43	0.15	1.33	-0.91, 3.57	0.24
cg16672562	0.31	-0.54, 1.15	0.48	-0.04	-0.64, 0.57	0.90

7.3.1.3 Bidirectional MR results

Bidirectional MR analysis was performed to explore causality in the cross-sectional association between cg27146050 methylation and BMI in adolescence (**Table 20**). The causal effect estimates were directionally consistent with those expected if BMI has causal effect on cg27146050 methylation but not with those expected if cg27146050 methylation as a causal effect on BMI. However, the confidence intervals were wide and spanned zero.

Table 20 – Results from bidirectional MR analysis of BMI and cg27146050 methylation in adolescence.

*Analyses are adjusted for bisulphite conversion batch only.

IV	Exposure	Outcome	Observed association*	Expected association	Difference between observed and expected estimates?
			β (95% CI)	β (95% CI)	p-value
cis-SNP score	methylation	log BMI	-0.0381 (-0.2937, 0.2176)	0.1027 (0.0315, 0.1739)	0.30
Standardised BMI allele score	log BMI	methylation	0.0014 (-0.0009, 0.0037)	0.0008 (0.0002, 0.0013)	0.55

7.3.2. BMI EWAS results

EWAS of BMI in ALSPAC at ages 7 and 15-17 years were performed. The sample size at age 7 was 913 children, and the sample size at age 15-17 was 784 adolescents.

No BMI-associated CpGs were identified with p-values below the epigenome-wide Bonferroni corrected threshold $p < 1.06 \times 10^{-7}$. Using a weaker suggestive p-value threshold, arbitrarily set at $p < 10^{-5}$, weak evidence was observed for associations between BMI and 8 CpGs at age 7 years and 6 CpGs at ages 15-17 years. EWAS results for these top associations between BMI and methylation are presented in **Table 21**, **Table 22**, **Figure 25** and **Figure 26**. Analyses were adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years.

After taking account of multiple testing (8 look-ups were performed in the adolescents' results; 6 look-ups were performed in the children's results), none of the BMI-CpG associations held in both childhood and adolescence. Out of the 14 CpGs, effect estimate directions were consistent across the two age groups for 9 CpGs: cg11836587, cg14965639, cg27205928, cg09797334, cg07285953, cg17820871, cg21698718, cg25110857 and cg27229251.

Table 21 – BMI EWAS results for BMI-CpG associations with $p < 10^{-5}$ in childhood.

Effect estimates are the change in methylation per 1kg/m^2 increase in BMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. Chr, chromosome.

			Age 7			Age 15-17		
CpG	**Chr**	**Gene**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
cg11836587	11		-6.65×10^{-3}	-9.12×10^{-3}, -4.17×10^{-3}	1.75×10^{-7}	-1.59×10^{-3}	-3.42×10^{-3}, 2.47×10^{-4}	0.09
cg22027865	8		5.90×10^{-3}	3.67×10^{-3}, 8.12×10^{-3}	2.60×10^{-7}	-3.28×10^{-4}	-1.88×10^{-3}, 1.23×10^{-3}	0.68
cg14965639	2	*STON1-GTF2A1L*	6.50×10^{-3}	3.91×10^{-3}, 9.09×10^{-3}	1.07×10^{-6}	1.08×10^{-3}	-8.89×10^{-4}, 3.04×10^{-3}	0.28
cg27205928	7	*C7orf50; MIR339*	4.65×10^{-3}	2.77×10^{-3}, 6.52×10^{-3}	1.45×10^{-6}	8.89×10^{-4}	-1.93×10^{-4}, 1.97×10^{-3}	0.11
cg09797334	11		-5.73×10^{-3}	-8.09×10^{-3}, -3.36×10^{-3}	2.45×10^{-6}	-1.87×10^{-3}	-3.54×10^{-3}, -2.06×10^{-4}	0.03
cg00244267	19	*SEMA6B*	2.72×10^{-3}	1.56×10^{-3}, 3.88×10^{-3}	4.64×10^{-6}	-3.63×10^{-4}	-1.15×10^{-3}, 4.29×10^{-4}	0.37
cg25693302	18	*NEDD4L*	2.91×10^{-4}	1.67×10^{-4}, 4.16×10^{-4}	5.41×10^{-6}	-6.19×10^{-5}	-1.57×10^{-4}, 3.34×10^{-5}	0.20
cg07285953	17	*CRYBA1*	3.75×10^{-3}	2.11×10^{-3}, 5.39×10^{-3}	8.18×10^{-6}	2.45×10^{-4}	-9.03×10^{-4}, 1.39×10^{-3}	0.68

Table 22 – BMI EWAS results for BMI-CpG associations with $p < 10^{-5}$ in adolescence.

Effect estimates are the change in methylation per 1kg/m^2 increase in BMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. Chr, chromosome.

			Age 15-17			Age 7		
CpG	**Chr**	**Gene**	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
cg00498089	6		2.77×10^{-3}	1.66×10^{-3}, 3.88×10^{-3}	1.34×10^{-6}	-8.01×10^{-4}	-2.65×10^{-3}, 1.05×10^{-3}	0.40
cg10220806	22	*DNAL4*	1.70×10^{-4}	9.97×10^{-5}, 2.40×10^{-4}	2.47×10^{-6}	-2.80×10^{-5}	-1.25×10^{-4}, 6.85×10^{-5}	0.57
cg17820871	12	*TESC*	1.78×10^{-3}	1.04×10^{-3}, 2.53×10^{-3}	3.27×10^{-6}	6.94×10^{-4}	-3.85×10^{-4}, 1.77×10^{-3}	0.21
cg21698718	17	*CCDC57*	2.29×10^{-3}	1.32×10^{-3}, 3.27×10^{-3}	4.55×10^{-6}	7.78×10^{-4}	-6.95×10^{-4}, 2.25×10^{-3}	0.30
cg25110857	2		1.60×10^{-3}	9.23×10^{-4}, 2.28×10^{-3}	4.58×10^{-6}	5.15×10^{-5}	-1.02×10^{-3}, 1.12×10^{-3}	0.92
cg27229251	7	*TTYH3*	2.80×10^{-3}	1.58×10^{-3}, 4.02×10^{-3}	7.98×10^{-6}	2.00×10^{-4}	-1.52×10^{-3}, 1.92×10^{-3}	0.82

Figure 25 - BMI EWAS results for BMI-CpG associations with $p < 10^{-5}$ in childhood.

Effect estimates are the change in methylation per 1kg/m² increase in BMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. Chr, chromosome.

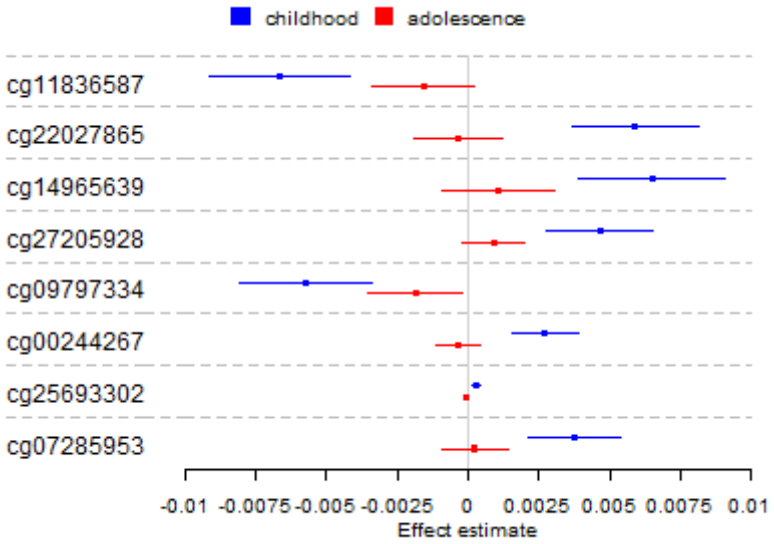
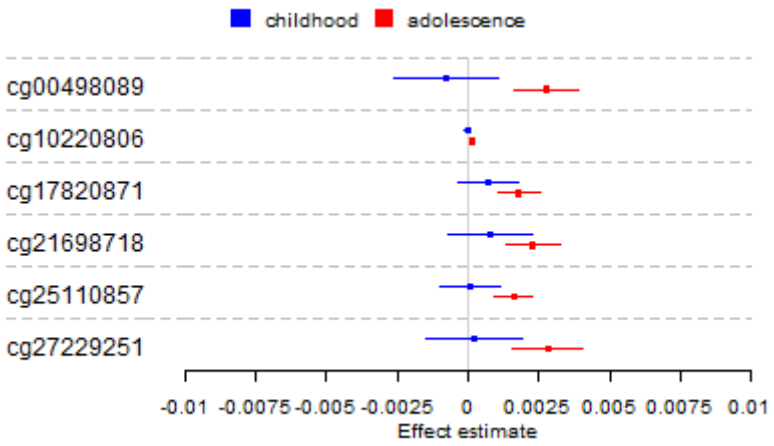


Figure 26 – BMI EWAS results for BMI-CpG associations with $p < 10^{-5}$ in adolescence.

Effect estimates are the change in methylation per 1kg/m² increase in BMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. Chr, chromosome.



7.3.3. Results from look-up of previously reported adult BMI CpGs in ALSPAC offspring

Wahl et al. and Mendelson et al. identified replicable epigenome-wide associations between BMI and 232 CpGs in adulthood.[94,95]

The association between BMI and the 232 CpGs was looked up in the results from the EWAS in ALSPAC in childhood and adolescence. 30 CpGs were associated at $p < 0.05$ (without correction for multiple testing) in childhood, and for 21 of these 30 CpGs the effect directions in childhood were consistent with those previously observed in adults (**Figure 27**). 33 CpGs were associated $p < 0.05$ in adolescence, of which 30 were directionally consistent with effects identified in adults (**Figure 28**).

The relationship between FMI and the 232 CpGs were also investigated. 22 CpGs were associated with FMI at $p < 0.05$ in childhood, of which 15 were directionally consistent with BMI-CpG associations in adults (**Figure 29**). 48 CpGs were associated with FMI at $p < 0.05$ in adolescence, and the effect directions for all 48 of these CpGs were consistent with the effect directions previously observed in adults (**Figure 30**).

Effect magnitudes observed in ALSPAC are not directly comparable with those observed by Wahl et al. and Mendelson et al. since they fitted different models or performed different transformations. Therefore only the effect directions between the studies were compared.

Figure 27 – BMI-associated CpGs previously identified in adults which also show an association of p<0.05 with BMI in childhood.

Effect estimates are the change in methylation per 1kg/m^2 increase in BMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. The +/- signs on the right-hand side of the plot represent the effect directions observed in adults by Wahl et al. and Mendelson et al. respectively; if they did not report an association this is represented by a "."

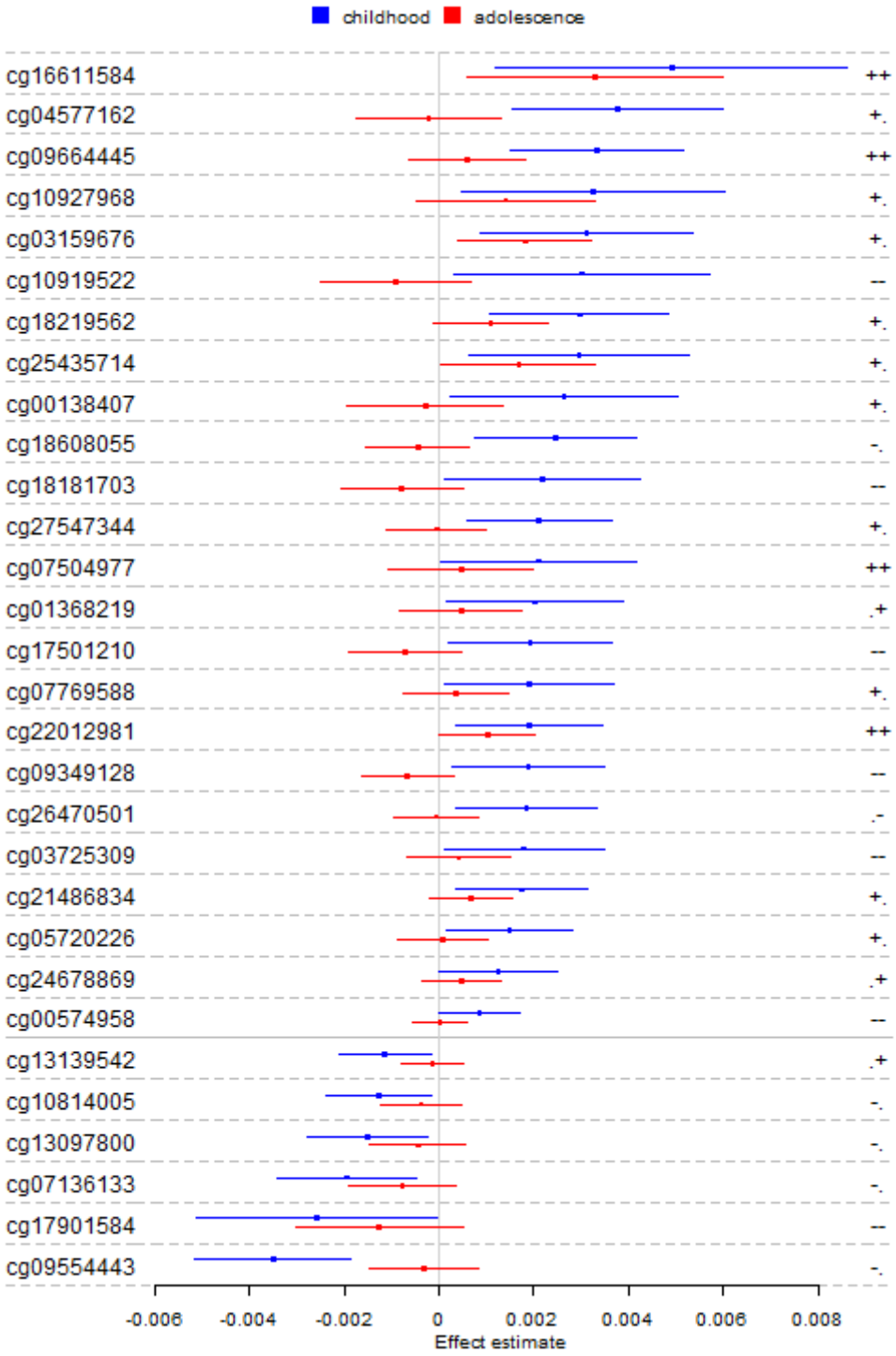


Figure 28 – BMI-associated CpGs previously identified in adults which also show an association of p<0.05 with BMI in adolescence.

Effect estimates are the change in methylation per 1kg/m^2 increase in BMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. The +/- signs on the right-hand side of the plot represent the effect directions observed in adults by Wahl et al. and Mendelson et al. respectively; if they did not report an association this is represented by a "."

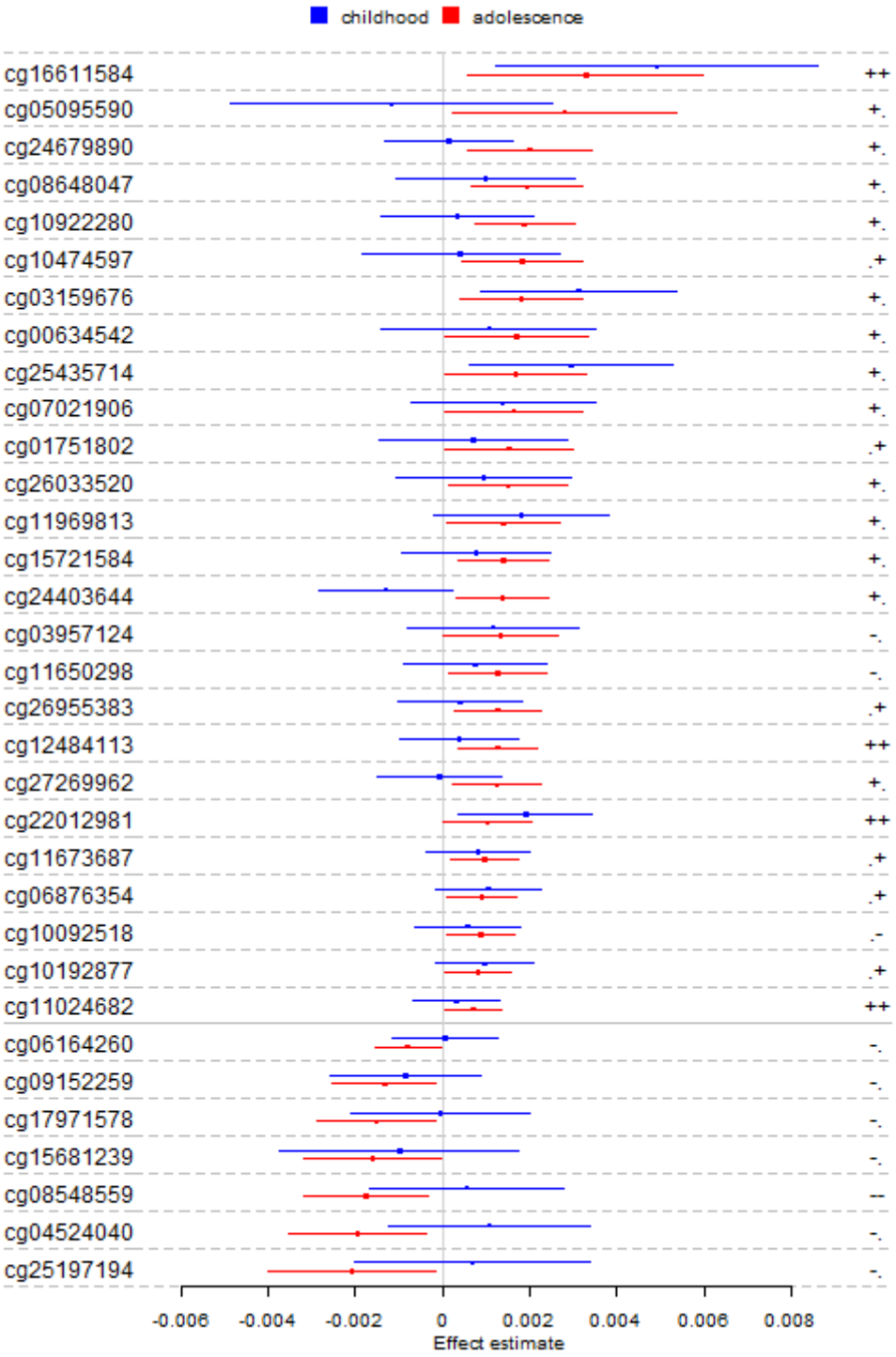


Figure 29 – BMI-associated CpGs previously identified in adults which also show an association of p<0.05 with FMI in childhood.

Effect estimates are the change in methylation per 1kg/m^2 increase in FMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. The +/- signs on the right-hand side of the plot represent the effect directions observed in adults by Wahl et al. and Mendelson et al. respectively; if they did not report an association this is represented by a "."

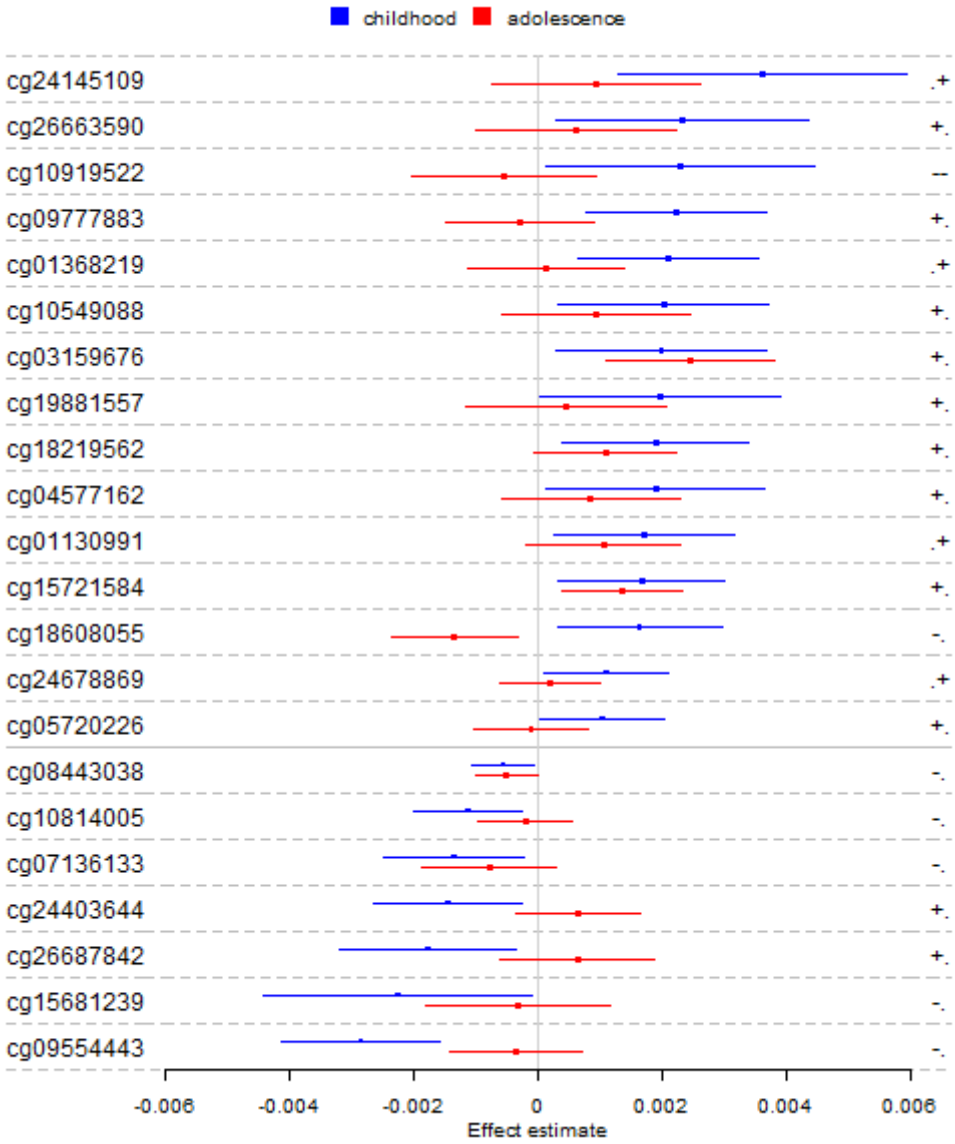
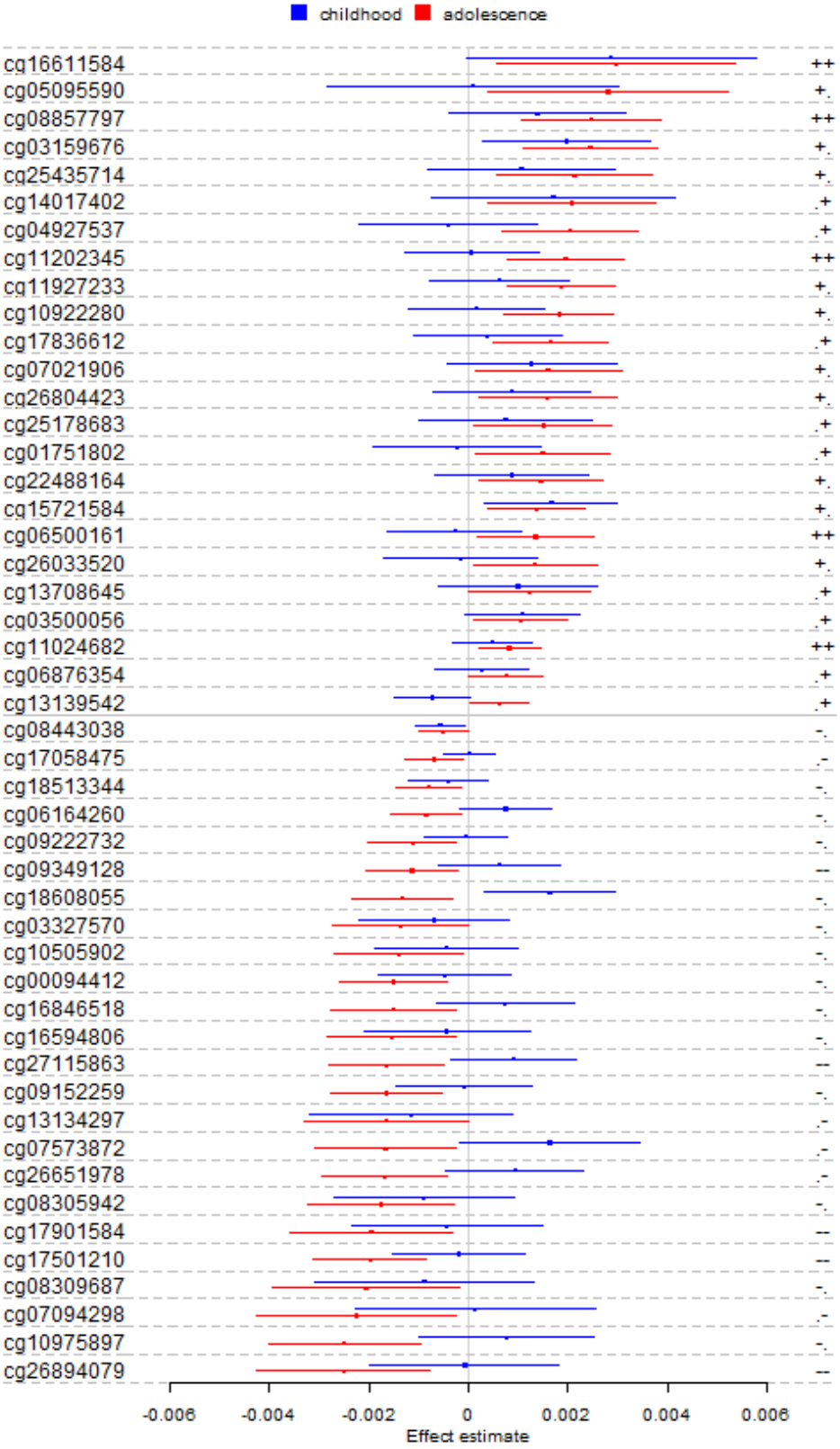


Figure 30 – BMI-associated CpGs previously identified in adolescence which also show an association of p<0.05 with FMI in childhood.

Effect estimates are the change in methylation per 1kg/m^2 increase in FMI. Models adjusted for age, sex, maternal education and cell counts at age 7 years, and additionally for smoking behaviour at age 15-17 years. The +/- signs on the right-hand side of the plot represent the effect directions observed in adults by Wahl et al. and Mendelson et al. respectively; if they did not report an association this is represented by a "."



7.3.4. Results from bidirectional MR

7.3.4.1 Results from BMI to methylation analyses

Weak associations have been identified between BMI and DNA methylation levels at 14 CpGs in the ALSPAC children and adolescents. To investigate whether BMI has a causal effect on these CpGs, the relationships between the GIANT BMI allele score and the 8 BMI-associated CpGs in childhood and 6 BMI-associated CpGs in adolescence were investigated. After taking account of multiple testing, none of the CpGs showed evidence of association with the allele score. The estimate directions of effect for 11 of the CpGs, however, were consistent with the effect estimate directions between those CpGs and BMI.

Table 23 – Associations between GIANT allele score at BMI-associated CpGs at age 7.

Effect estimates are the change in methylation per unit increase in BMI allele score. Models adjusted for age, sex, maternal education and cell counts.

CpG	Chr	Gene	Beta	95% CI	p-value	Effect direction consistent with BMI-CpG effect direction?
cg11836587	11		1.18×10^{-4}	-5.55×10^{-4}, 7.92×10^{-4}	0.731	No
cg22027865	8		2.01×10^{-4}	-4.49×10^{-4}, 8.51×10^{-4}	0.545	Yes
cg14965639	2	STON1-GTF2A1L	6.80×10^{-5}	-6.27×10^{-4}, 7.63×10^{-4}	0.848	Yes
cg27205928	7	C7orf50; MIR339	3.15×10^{-4}	-2.28×10^{-4}, 8.57×10^{-4}	0.256	Yes
cg09797334	11		-2.59×10^{-4}	-9.25×10^{-4}, 4.07×10^{-4}	0.446	Yes
cg00244267	19	SEMA6B	1.23×10^{-4}	-2.25×10^{-4}, 4.70×10^{-4}	0.490	Yes
cg25693302	18	NEDD4L	-1.87×10^{-6}	-3.84×10^{-5}, 3.46×10^{-5}	0.920	No
cg07285953	17	CRYBA1	5.59×10^{-4}	8.34×10^{-5}, 1.03×10^{-3}	0.022	Yes

Table 24 – Associations between GIANT allele score at BMI-associated CpGs at age 15-17.

Effect estimates are the change in methylation per unit increase in BMI allele score. Models adjusted for age, sex, maternal education, cell counts and smoking behaviour.

CpG	Chr	Gene	Beta	95% CI	p-value	Effect direction consistent with BMI-CpG effect direction?
cg00498089	**6**		2.14×10^{-4}	-3.16×10^{-4}, 7.44×10^{-4}	0.429	Yes
cg10220806	22	DNAL4	3.08×10^{-5}	-3.60×10^{-6}, 6.52×10^{-5}	0.080	Yes
cg17820871	12	TESC	-1.19×10^{-4}	-5.02×10^{-4}, 2.65×10^{-4}	0.544	No
cg21698718	17	CCDC57	3.13×10^{-4}	-1.50×10^{-4}, 7.75×10^{-4}	0.186	Yes
cg25110857	2		3.41×10^{-4}	1.36×10^{-5}, 6.68×10^{-4}	0.042	Yes
cg27229251	7	TTYH3	2.91×10^{-5}	-5.77×10^{-4}, 6.35×10^{-4}	0.925	Yes

7.3.4.2 Results from methylation to BMI MR

Only two of the 14 CpGs (cg11836587 and cg14965639) had mQTL available, and hence MR analysis was only performed for these two CpGs. Results from the 2-sample MR analysis are presented below in **Table 25**. These results do not provide evidence for causal effects of these CpGs on BMI.

Table 25 – Results from 2-sample MR analysis of the effect of BMI-associated CpGs on BMI.

CpG	SNP	Beta	95% CI	p-value	Effect direction consistent with BMI-CpG effect direction?
cg11836587	rs7938259	0.014	-0.007, 0.035	0.188	No
cg14965639	rs17397707	0.003	-0.016, 0.022	0.747	Yes

7.3.5. Results from look-up of age 7 BMI-associated CpGs with diet PCs

The relationship between dietary behaviour and the 8 BMI-associated CpGs at age 7 years was investigated, and the results are presented in **Table 26**. Analyses were adjusted for age, sex, maternal education and estimated cell counts. No robust evidence of association between the CpGs and diet PCs was observed.

Table 26 – Relationship between dietary behaviour and BMI-associated CpGs at age 7 years.

	"health aware" PC			"packed lunch" PC		
CpG	**Beta**	**95% CI**	**p-value**	**Beta**	**95% CI**	**p-value**
cg11836587	-2.75×10^{-4}	-3.03×10^{-3}, 2.48×10^{-3}	0.845	-9.56×10^{-4}	-4.01×10^{-3}, 2.10×10^{-3}	0.539
cg22027865	-1.74×10^{-4}	-2.60×10^{-3}, 2.25×10^{-3}	0.888	-6.89×10^{-4}	-3.43×10^{-3}, 2.05×10^{-3}	0.623
cg14965639	1.30×10^{-3}	-1.51×10^{-3}, 4.11×10^{-3}	0.364	1.28×10^{-3}	-1.80×10^{-3}, 4.36×10^{-3}	0.415
cg27205928	2.16×10^{-4}	-1.88×10^{-3}, 2.31×10^{-3}	0.839	1.47×10^{-3}	-8.58×10^{-4}, 3.81×10^{-3}	0.216
cg09797334	-1.60×10^{-4}	-2.72×10^{-3}, 2.40×10^{-3}	0.902	-3.20×10^{-3}	-6.04×10^{-3}, -3.64×10^{-4}	0.027
cg00244267	-3.05×10^{-4}	-1.65×10^{-3}, 1.04×10^{-3}	0.658	-9.75×10^{-4}	-2.44×10^{-3}, 4.93×10^{-4}	0.193
cg25693302	9.13×10^{-5}	-4.84×10^{-5}, 2.31×10^{-4}	0.201	-5.83×10^{-6}	-1.61×10^{-4}, 1.50×10^{-4}	0.941
cg07285953	-4.50×10^{-4}	-2.25×10^{-3}, 1.35×10^{-3}	0.625	5.07×10^{-4}	-1.49×10^{-3}, 2.51×10^{-3}	0.620

7.4. Discussion

HIF3A methylation and BMI in childhood and adolescence

Analyses were performed to investigate whether associations between *HIF3A* methylation and BMI, previously identified in a separate study of adults,[91] are also present in childhood and adolescence in ALSPAC. Cross-sectional analysis observed an association between methylation at cg27146050 and BMI in adolescence but not childhood. This association was of a similar magnitude to that previously observed in adults.[91] The cross-sectional analysis of cg27146050 and BMI in adolescence was repeated with adjustment for estimated cell composition, and the results of this analysis did not differ from results from the original model without cell-type correction.[109] Results from the longitudinal and MR analyses in ALSPAC suggest that, in a causal relationship between BMI and *HIF3A* methylation, the direction of effect is more likely to be from BMI to *HIF3A* methylation than the reverse direction. The full manuscript of this work can be found in **Appendix B** of this thesis.

Epigenome-wide analyses of BMI and methylation

These analyses investigated the relationship between BMI and methylation in childhood and adolescence. None of the BMI-CpG associations reached epigenome-wide significance, but 14 CpGs showed suggestive evidence of association with BMI with $p < 10^{-5}$. Of these 14 associations, 8 were identified in childhood and 6 were identified in adolescence.

The available sample sizes for this chapter's analyses in childhood (n=913) and adolescence (n=784) are far smaller than those used in the EWAS discovery analyses by Wahl et al. (n=5,387) and Mendelson et al. (n=3,743).[94,95] This may partly explain why the above analyses in ALSPAC were not able to detect robust associations between BMI and methylation, though a more recent smaller BMI EWAS in 374 pre-school children was able to detect several BMI-associated CpGs.[232]

232 previously identified BMI-CpG associations in adults were followed up in ALSPAC in childhood and adolescence. 30 of those CpGs showed some evidence of association in

childhood and 33 of those CpGs showed some evidence of association in adolescence. Most of the effect estimates for these associations were consistent in effect direction with those in adults. Associations in adolescence tended to be more directionally consistent with the adult associations than the childhood associations were with the adult associations (30 out of 33 CpGs in adolescence compared to 21 out of 30 CpGs in childhood). Four of the CpGs showed some evidence of association in both childhood and adolescence – these were cg16611584 (nearest gene *AKAP10*), cg03159676 (nearest gene *GSE1*), cg25435714 (nearest gene *RN7SL142P*) and cg22012981 (nearest gene *ACOX2*).

GWAS have identified links for *AKAP10* with mean platelet volume (MPV), platelet count and reticulocyte count, and *GSE1* with platelet count.[233,234] A literature review of the relationship between MPV and cardiovascular diseases found several links, including links with obesity, diabetes and myocardial infarction.[235] Thus a plausible biological pathway may link BMI with methylation variation at the *AKAP10* locus and subsequent cardiovascular disease, although this would require much more detailed investigation before concrete inferences could be made.

Causal inference analyses were performed to investigate whether any of the suggestively BMI-associated CpGs had a causal effect on BMI, or vice versa. These analyses provided little evidence for or against causal effects in either direction. Associations between the GIANT BMI allele score and BMI-associated CpGs were weak/lacking but mainly consistent in direction with the BMI-CpG effect estimates. Suitable genetic instruments to assess the causal effect of methylation on BMI were only available for two CpGs, and, after removing SNPs in LD, only one mQTL SNP was available for each of those CpGs. The capacity to execute MR using mQTLs as instrumental variables for site specific DNA methylation will improve as the catalogue of available mQTLs increases. Large scale GWAS of methylation are underway which should make this type of analysis more tractable in the near future.

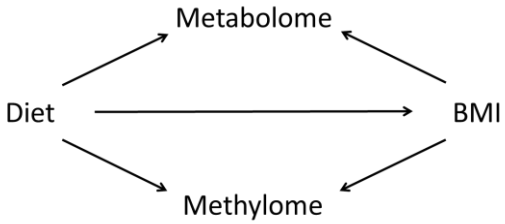
Analyses investigating the relationship between BMI-associated diet PCs and suggestively BMI-associated CpGs did not identify any associations. Since diet is a

complex phenotype and the CpGs were not robustly associated with BMI, this lack of identifiable associations between the diet PCs and these CpGs is not surprising and does not rule out the possibility that diet may indeed play a role in the relationship between BMI and methylation.

A limitation of the analyses in this chapter is that the only tissue studied is blood. However, when studying associations with BMI, it may be more appropriate to study adipose tissue since methylation patterns are tissue-specific.[81,109] BMI-CpG associations may be greater in adipose tissue, as was found to be the case for *HIF3A* methylation where the observed percentage change in BMI per 0.1 unit increase in methylation ß-value was greater in adipose tissue than blood.[91]

CHAPTER 8. DISCUSSION

This thesis explored the role of the metabolome and the methylome in the relationship between dietary behaviour and BMI. Analyses conducted to achieve this aim included a GWAS to identify genetic instruments for diet; analyses to assess the diet-BMI relationship; analyses to identify aspects of the metabolome and methylation that may mediate the diet-BMI relationship; and MR and mediation analyses to investigate these potential mediators.



8.1. Genetic determinants of dietary intake

Main findings

The GWAS of macronutrient intake identified five diet-SNP associations that replicated within UK Biobank. These included an association between rs516246 and polyunsaturated fat intake, which was novel in this GWAS, and an association between rs838133 and protein intake, which was identified in a previous diet GWAS.[134] These two SNPs are in nearby genes – rs516246 is in *FUT2*, and rs838133 is in *FGF21*. The LD between rs516246 and rs838133 is R^2=0.364, hence they may be tagging the same causal effect. *FUT2* influences secretor status and intestinal microbiota composition,[181,182] and previous GWAS have linked *FUT2* SNPs to Crohn's disease,[158-161] cholesterol,[157] vitamin B12 levels,[178-180] folate pathway vitamin levels[164,165] and serum lipase activity.[174]

Strengths and limitations

The macronutrient intake GWAS in UK Biobank was carried out in a sample twice the size of any previous diet GWAS (~144,000 people, compared to the previous largest diet GWAS sample of ~71,000 people in the CHARGE and DietGen consortia[134,140]). UK Biobank has a panel of ~11 million SNPs, compared to ~2.6 million SNPs in previous

macronutrient GWAS elsewhere.[134,140] This large panel of SNPs is needed since macronutrient intake appears to be a highly heterogeneous trait influenced by many small genetic effects.

A major limitation in studies of dietary behaviour is measurement error for diet. This is an issue in UK Biobank, though slightly less so since dietary intake was measured multiple times and averaged across the various repeat measures, which hopefully reduced measurement error through reducing the impact of daily variation on the data.

Future directions

Conducting a diet GWAS with an even larger sample may identify further diet-SNP associations, however I think it would be more fruitful to focus effort on using dietary measures that are less reliant on the (often biased) self-report of dietary intake, for example cameras and software to digitally capture and estimate food groups and portion size. Alternatively, more objective measures of dietary constituents, such as metabolites, could be used.

8.2. Implementing MR to understand diet-BMI relationship

Main findings

After observing strong links between dietary behaviour and BMI, these links were then explored within a causal inference framework. Two-sample MR analyses, conducted to estimate the causal effect of macronutrient intake on BMI, observed weak evidence suggesting that increases in protein and polyunsaturated fat intake lead to decreases in BMI, whereas increases in fat and saturated fat intake lead to increases in BMI.

The 97-SNP BMI allele score from Locke et al.[52] is not suitable for use in an MR framework to assess the causal effect of BMI on dietary behaviour since some of the loci are thought to influence appetite,[208,213] and this violates the MR requirement for the outcome variable to only be associated with the genetic instrument through the

exposure variable.[55] Instead, leave-one-out (LOO) weighted allele scores were created for each of 25 functional categories (**4.2.1.3**) to explore whether associations between dietary behaviour and the BMI allele scores were driven by a single aspect of the allele score, for example SNPs in the hypothalamic expression and regulatory function category. Dietary behaviour was captured using macronutrient intake in UK Biobank and the diet PCs in ALSPAC. The associations between the LOO allele scores and dietary behaviour were weaker for allele scores without SNPs in the neuronal developmental processes category and the hypothalamic expression and regulatory function category, which suggests that some BMI SNPs may exert their effect on BMI through dietary behaviour.

Strengths and limitations

A strength of these analyses is the implementation of a two-sample MR framework as this enabled use of GWAS summary data from a large published BMI GWAS.[52] The macronutrient genetic instruments used in these causal analyses were identified in UK Biobank, hence one-sample MR could not be conducted in UK Biobank as this would have led to overfitting.[210,211] However, using MR to explore the causal effect of diet on BMI was challenging because few diet-associated SNPs have been identified. For each macronutrient no more than two SNPs were available as genetic instruments for use in MR analysis of macronutrient intake on BMI.

Conducting causal analyses to explore the effect of BMI on diet also proved challenging. Some BMI SNPs are thought to influence BMI through diet-related traits such as appetite regulation, and hence these SNPs do not meet the requirement in MR for the instrument to be independent of the outcome given the exposure.[55]

Future directions

If better genetic instruments for dietary behaviour are available for future analyses, this should help to the clarify causal relationship between diet and BMI, and more specifically which components of diet are most detrimental in causing accrual of body weight. A more detailed knowledge of the biological functions of SNPs in the BMI allele

score would hopefully identify some SNPs that, given BMI, are independent of dietary behaviour.

8.3. Dietary and BMI influences on the metabolome

Main findings

Strong links were observed between diet, BMI and the metabolome. Cross-sectional associations were observed in ALSPAC in childhood and adolescence between BMI and several metabolites, including VLDL and HDL concentration measures, apolipoproteins, fatty acids and amino acids. Effect directions were mostly consistent with those previously observed in adults, indicating that the link between BMI and the metabolome starts in childhood.[70,71,74] Results from MR analyses of BMI on metabolites in the ALSPAC children and adolescents suggest that BMI has a causal effect on several metabolites. Previous studies in adults have also observed evidence indicating a causal effect of BMI on the metabolome.[70,75,76] Several BMI-associated metabolites were also associated with dietary behaviour in the ALSPAC children, including branched-chain amino acids.

Strengths and limitations

Previous studies have identified strong links between adiposity and the metabolome, however these studies were mostly conducted in adults.[68-78] Two studies were carried out in children, but the sample sizes were small (n<250) and neither investigated causality.[68,69] In contrast, the analyses of adiposity and the metabolome in this thesis were conducted in far greater sample sizes (n=5414 in childhood and n=3286 in adolescence).

Future directions

Further analyses are needed to clarify the directions of causality between BMI and metabolites and between dietary behaviour and the metabolites. Although this study was larger than previous studies of children, even bigger studies are needed for the power to detect smaller effects.

8.4. Direction of causal pathways between BMI and DNA methylation

Main findings

Several associations between BMI and methylation, previously identified in adult studies elsewhere, are also present in childhood and adolescence in ALSPAC. The associations in ALSPAC tended to be weaker, probably due to using a smaller sample size than the previous studies in adults.[91,94,95]

Causal inference analyses were unable to, with any degree of certainty, identify or rule out a causal effect of BMI on methylation or methylation on BMI. However, results from longitudinal and MR analyses suggest that the direction of effect between BMI and *HIF3A* methylation is more likely to be from BMI to *HIF3A* methylation than the reverse direction.

Strengths and limitations

EWAS are susceptible to confounding since the epigenetic patterns vary across the life course, and epigenetic variation can be a cause or a consequence of trait. BMI EWAS conducted in ALSPAC lacked power – larger samples will be required to identify epigenome-wide significant BMI-associated CpGs.

Future directions

Further EWAS using larger samples are needed to identify BMI-associated CpGs in children and adolescents. BMI-associated methylation variation has been applied to predict downstream consequences of BMI including type 2 diabetes,[94] so although the association studies to date have shed little light on the causal role of DNA methylation in this context, there is merit in utilising methylation as a predictor of future comorbidities. The use of methylation variation in children as a predictor of later adverse health outcomes warrants further study.

8.5. Implementing MR in molecular mediation

Main findings

Analyses in the ALSPAC children found that several BMI-associated metabolites were also associated with diet. In contrast, no associations were identified between BMI-associated CpGs and dietary behaviour.

The relationships between diet, BMI and the metabolites suggest that five of the metabolites (HDL particle size, MUFA, creatinine, and very large HDL cholesterol and cholesterol esters) may be potential mediators in the relationship between diet and BMI (or that the relationship between diet and those metabolites may be mediated by BMI). 2-sample MR analyses performed to investigate the causal relationships between these metabolites and BMI observed evidence suggesting that BMI has a causal effect on HDL particle size and very large HDL cholesterol and cholesterol esters. Results from single sample MR analysis conducted in the ALSPAC children suggest that BMI also has a causal effect on creatine levels. 2-sample MR analyses were unable to detect any causal effects of the metabolites on BMI, perhaps due to a lack of power.

Bringing together findings from the MR analyses and mediation analyses, the results suggest that BMI mediates the effect of dietary behaviour on very large HDL cholesterol and cholesterol esters, HDL particle size and creatinine.

Strengths and limitations

A limitation of these analyses is that the causal effect of diet on BMI and metabolites was assumed but not assessed as there are no known genetic instruments for the diet PCs (**4.2.2.2**). PCs are unique to the dataset from which they are generated, and hence comparing PCs from one study with those from different study (or the same study at a different timepoint) is not straightforward. However, when dealing with complex dietary data, PCA is helpful for the identification of dietary behaviours that are worthy of more detailed scrutiny.

Future directions

The relationships between diet, BMI and the metabolome and methylome are complex and warrant further investigation. Larger sample sizes and more refined measures of dietary behaviour and adiposity are needed.

8.6. Main conclusions

The work presented here confirms that a healthy diet is important in reducing obesity and identifies some potential molecular mechanisms by which this may occur. Importantly, my results indicate that the molecular mechanisms underpinning this relationship become established in childhood, emphasising the importance of early intervention to reduce risk of later cardiometabolic disease. Furthermore, the data presented in this thesis suggest that obesity influences molecular profiles, in particular DNA methylation. These changes may be informative in predicting adverse consequences of obesity, potentially for stratifying groups of individuals for targeted intervention.

This thesis helps to improve our understanding of the relationship between molecular intermediates and obesity. Molecular intermediates such as the metabolome and methylome provide additional opportunities for intervention, with the aim of either the prevention and treatment of obesity itself, or of its comorbidities.

My results emphasise the importance of a healthy lifestyle from an early age. This supports current public health strategies and augments the evidence for the importance of childhood lifestyle and dietary interventions.

References

1. Guh DP, Zhang W, Bansback N, Amarsi Z, Birmingham CL, Anis AH. The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC Public Health* 2009; **9**: 88.
2. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. *World Health Organ Tech Rep Ser* 2000; **894**: i-xii, 1-253.
3. Grover SA, Kaouache M, Rempel P, et al. Years of life lost and healthy life-years lost from diabetes and cardiovascular disease in overweight and obese people: a modelling study. *Lancet Diabetes Endocrinol* 2015; **3**(2): 114-22.
4. Wang YC, McPherson K, Marsh T, Gortmaker SL, Brown M. Health and economic burden of the projected obesity trends in the USA and the UK. *Lancet* 2011; **378**(9793): 815-25.
5. Youngson NA, Morris MJ. What obesity research tells us about epigenetic mechanisms. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2013; **368**(1609): 20110337.
6. Riddoch CJ, Leary SD, Ness AR, et al. Prospective associations between objective measures of physical activity and fat mass in 12-14 year old children: the Avon Longitudinal Study of Parents and Children (ALSPAC). *Bmj* 2009; **339**: b4544.
7. Hills AP, Andersen LB, Byrne NM. Physical activity and obesity in children. *British journal of sports medicine* 2011; **45**(11): 866-70.
8. Singh AS, Mulder C, Twisk JW, van Mechelen W, Chinapaw MJ. Tracking of childhood overweight into adulthood: a systematic review of the literature. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2008; **9**(5): 474-88.
9. Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. *Circulation* 2009; **119**(24): 3078-84.
10. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. *Journal of the American College of Cardiology* 2009; **54**(14): 1209-27.
11. Steinberger J, Daniels SR, American Heart Association Atherosclerosis H, Obesity in the Young C, American Heart Association Diabetes C. Obesity, insulin resistance, diabetes, and cardiovascular risk in children: an American Heart Association scientific statement from the Atherosclerosis, Hypertension, and Obesity in the Young Committee (Council on Cardiovascular Disease in the Young) and the Diabetes Committee (Council on Nutrition, Physical Activity, and Metabolism). *Circulation* 2003; **107**(10): 1448-53.
12. Strong JP, Malcom GT, McMahan CA, et al. Prevalence and extent of atherosclerosis in adolescents and young adults: implications for prevention from the Pathobiological Determinants of Atherosclerosis in Youth Study. *JAMA : the journal of the American Medical Association* 1999; **281**(8): 727-35.
13. Berenson GS, Srinivasan SR, Bao W, Newman WP, 3rd, Tracy RE, Wattigney WA. Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults. The Bogalusa Heart Study. *The New England journal of medicine* 1998; **338**(23): 1650-6.
14. Brown T, Kelly S, Summerbell C. Prevention of obesity: a review of interventions. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2007; **8 Suppl 1**: 127-30.
15. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* 2015; **12**(3): e1001779.
16. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics* 2014.

17. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. *Human molecular genetics* 2016; **25**(R2): R149-R56.
18. Perez-Escamilla R, Obbagy JE, Altman JM, et al. Dietary energy density and body weight in adults and children: a systematic review. *J Acad Nutr Diet* 2012; **112**(5): 671-84.
19. Tucker LA, Seljaas GT, Hager RL. Body fat percentage of children varies according to their diet composition. *J Am Diet Assoc* 1997; **97**(9): 981-6.
20. Smith AD, Emmett PM, Newby PK, Northstone K. Dietary patterns and changes in body composition in children between 9 and 11 years. *Food Nutr Res* 2014; **58**.
21. Fraser LK, Edwards KL, Cade JE, Clarke GP. Fast food, other food choices and body mass index in teenagers in the United Kingdom (ALSPAC): a structural equation modelling approach. *International journal of obesity* 2011; **35**(10): 1325-30.
22. Satia-Abouta J, Patterson RE, Schiller RN, Kristal AR. Energy from fat is associated with obesity in U.S. men: results from the Prostate Cancer Prevention Trial. *Preventive medicine* 2002; **34**(5): 493-501.
23. Schulze MB, Fung TT, Manson JE, Willett WC, Hu FB. Dietary patterns and changes in body weight in women. *Obesity* 2006; **14**(8): 1444-53.
24. Fogelholm M, Anderssen S, Gunnarsdottir I, Lahti-Koski M. Dietary macronutrients and food consumption as determinants of long-term weight change in adult populations: a systematic literature review. *Food Nutr Res* 2012; **56**.
25. Newby PK, Weismayer C, Akesson A, Tucker KL, Wolk A. Longitudinal changes in food patterns predict changes in weight and body mass index and the effects are greatest in obese women. *The Journal of nutrition* 2006; **136**(10): 2580-7.
26. Pachucki MA. Food pattern analysis over time: unhealthful eating trajectories predict obesity. *International journal of obesity* 2012; **36**(5): 686-94.
27. Northstone K, Emmett P. Multivariate analysis of diet in children at four and seven years of age and associations with socio-demographic characteristics. *European journal of clinical nutrition* 2005; **59**(6): 751-60.
28. Karnehed N, Tynelius P, Heitmann BL, Rasmussen F. Physical activity, diet and gene-environment interactions in relation to body mass index and waist circumference: the Swedish young male twins study. *Public health nutrition* 2006; **9**(7): 851-8.
29. Craig LC, McNeill G, Macdiarmid JI, Masson LF, Holmes BA. Dietary patterns of school-age children in Scotland: association with socio-economic indicators, physical activity and obesity. *The British journal of nutrition* 2010; **103**(3): 319-34.
30. Berkey CS, Rockett HR, Field AE, et al. Activity, dietary intake, and weight changes in a longitudinal study of preadolescent and adolescent boys and girls. *Pediatrics* 2000; **105**(4): E56.
31. Northstone K, Smith AD, Newby PK, Emmett PM. Longitudinal comparisons of dietary patterns derived by cluster analysis in 7- to 13-year-old children. *The British journal of nutrition* 2013; **109**(11): 2050-8.
32. Ritchie LD, Spector P, Stevens MJ, et al. Dietary patterns in adolescence are related to adiposity in young adulthood in black and white females. *The Journal of nutrition* 2007; **137**(2): 399-406.
33. Basterfield L, Jones AR, Parkinson KN, et al. Physical activity, diet and BMI in children aged 6-8 years: a cross-sectional analysis. *BMJ Open* 2014; **4**(6): e005001.
34. Elliott SA, Truby H, Lee A, Harper C, Abbott RA, Davies PS. Associations of body mass index and waist circumference with: energy intake and percentage energy from macronutrients, in a cohort of Australian children. *Nutrition journal* 2011; **10**: 58.
35. Anderson JJ, Celis-Morales CA, Mackay DF, et al. Adiposity among 132 479 UK Biobank participants; contribution of sugar intake vs other macronutrients. *International journal of epidemiology* 2017; **46**(2): 492-501.

36. Laska MN, Murray DM, Lytle LA, Harnack LJ. Longitudinal associations between key dietary behaviors and weight gain over time: transitions through the adolescent years. *Obesity* 2012; **20**(1): 118-25.
37. Thompson FE, Subar AF. Dietary assessment methodology. In: Coulston AM, Boushey CJ, Ferruzzi M, Delahanty L, eds. Nutrition in the Prevention and Treatment of Disease. 4th ed: Elsevier Science; 2017: 5-48.
38. Anderson EL, Tilling K, Fraser A, et al. Estimating trajectories of energy intake through childhood and adolescence using linear-spline multilevel models. *Epidemiology* 2013; **24**(4): 507-15.
39. Bingham SA, Gill C, Welch A, et al. Validation of dietary assessment methods in the UK arm of EPIC using weighed records, and 24-hour urinary nitrogen and potassium and serum vitamin C and carotenoids as biomarkers. *International journal of epidemiology* 1997; **26 Suppl 1**: S137-51.
40. Rebro SM, Patterson RE, Kristal AR, Cheney CL. The effect of keeping food records on eating patterns. *J Am Diet Assoc* 1998; **98**(10): 1163-5.
41. Livingstone MB, Black AE. Markers of the validity of reported energy intake. *The Journal of nutrition* 2003; **133 Suppl 3**: 895S-920S.
42. Hebert JR, Clemow L, Pbert L, Ockene IS, Ockene JK. Social desirability bias in dietary self-report may compromise the validity of dietary intake measures. *International journal of epidemiology* 1995; **24**(2): 389-98.
43. Miller TM, Abdel-Maksoud MF, Crane LA, Marcus AC, Byers TE. Effects of social approval bias on self-reported fruit and vegetable consumption: a randomized controlled trial. *Nutrition journal* 2008; **7**: 18.
44. Smith AD, Emmett PM, Newby PK, Northstone K. A comparison of dietary patterns derived by cluster and principal components analysis in a UK cohort of children. *European journal of clinical nutrition* 2011; **65**(10): 1102-9.
45. Bowman SA, Gortmaker SL, Ebbeling CB, Pereira MA, Ludwig DS. Effects of fast-food consumption on energy intake and diet quality among children in a national household survey. *Pediatrics* 2004; **113**(1 Pt 1): 112-8.
46. Rankinen T, Bouchard C. Genetics of food intake and eating behavior phenotypes in humans. *Annu Rev Nutr* 2006; **26**: 413-34.
47. Must A, Anderson SE. Body mass index in children and adolescents: considerations for population-based applications. *International journal of obesity* 2006; **30**(4): 590-4.
48. Reilly JJ. Diagnostic accuracy of the BMI for age in paediatrics. *International journal of obesity* 2006; **30**(4): 595-7.
49. Okorodudu DO, Jumean MF, Montori VM, et al. Diagnostic performance of body mass index to identify obesity as defined by body adiposity: a systematic review and meta-analysis. *International journal of obesity* 2010; **34**(5): 791-9.
50. Lawlor DA, Benfield L, Logue J, et al. Association between general and central adiposity in childhood, and change in these, with cardiovascular risk factors in adolescence: prospective cohort study. *Bmj* 2010; **341**: c6224.
51. Elks CE, den Hoed M, Zhao JH, et al. Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol (Lausanne)* 2012; **3**: 29.
52. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518**(7538): 197-206.
53. Yengo L, Sidorenko J, Kemper KE, et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Human molecular genetics* 2018; **27**(20): 3641-9.
54. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology* 2003; **32**(1): 1-22.

55. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* 2008; **27**(8): 1133-63.
56. Dunn WB, Ellis DI. Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry* 2005; **24**(4): 285-94.
57. Soininen P, Kangas AJ, Wurtz P, Suna T, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* 2015; **8**(1): 192-206.
58. Kettunen J, Tukiainen T, Sarin AP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* 2012; **44**(3): 269-76.
59. Bouchard-Mercier A, Rudkowska I, Lemieux S, Couture P, Vohl MC. The metabolic signature associated with the Western dietary pattern: a cross-sectional study. *Nutrition journal* 2013; **12**: 158.
60. Floegel A, von Ruesten A, Drogan D, et al. Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam. *European journal of clinical nutrition* 2013; **67**(10): 1100-8.
61. Altmaier E, Kastenmuller G, Romisch-Margl W, et al. Questionnaire-based self-reported nutrition habits associate with serum metabolism as revealed by quantitative targeted metabolomics. *European journal of epidemiology* 2011; **26**(2): 145-56.
62. Socha P, Grote V, Gruszfeld D, et al. Milk protein intake, the metabolic-endocrine response, and growth in infancy: data from a randomized clinical trial. *The American journal of clinical nutrition* 2011; **94**(6 Suppl): 1776S-84S.
63. Menni C, Zhai G, Macgregor A, et al. Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. *Metabolomics* 2013; **9**(2): 506-14.
64. Playdon MC, Moore SC, Derkach A, et al. Identifying biomarkers of dietary patterns by using metabolomics. *The American journal of clinical nutrition* 2017; **105**(2): 450-65.
65. Pallister T, Jennings A, Mohney RP, et al. Characterizing Blood Metabolomics Profiles Associated with Self-Reported Food Intakes in Female Twins. *PloS one* 2016; **11**(6): e0158568.
66. Ismail NA, Posma JM, Frost G, Holmes E, Garcia-Perez I. The role of metabonomics as a tool for augmenting nutritional information in epidemiological studies. *Electrophoresis* 2013; **34**(19): 2776-86.
67. Guertin KA, Moore SC, Sampson JN, et al. Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *The American journal of clinical nutrition* 2014; **100**(1): 208-17.
68. Wahl S, Yu Z, Kleber M, et al. Childhood obesity is associated with changes in the serum metabolite profile. *Obes Facts* 2012; **5**(5): 660-70.
69. Perng W, Gillman MW, Fleisch AF, et al. Metabolomic profiles and childhood obesity. *Obesity* 2014; **22**(12): 2570-8.
70. Wurtz P, Wang Q, Kangas AJ, et al. Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS medicine* 2014; **11**(12): e1001765.
71. Bogl LH, Kaye SM, Ramo JT, et al. Abdominal obesity and circulating metabolites: A twin study approach. *Metabolism* 2016; **65**(3): 111-21.
72. Boulet MM, Chevrier G, Grenier-Larouche T, et al. Alterations of plasma metabolite profiles related to adipose tissue distribution and cardiometabolic risk. *Am J Physiol Endocrinol Metab* 2015; **309**(8): E736-46.
73. Ho JE, Larson MG, Ghorbani A, et al. Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes. *PloS one* 2016; **11**(2): e0148361.
74. Moore SC, Matthews CE, Sampson JN, et al. Human metabolic correlates of body mass index. *Metabolomics* 2014; **10**(2): 259-69.

75. Holmes MV, Lange LA, Palmer T, et al. Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. *Am J Hum Genet* 2014; **94**(2): 198-208.
76. Freathy RM, Timpson NJ, Lawlor DA, et al. Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes* 2008; **57**(5): 1419-26.
77. Chen HH, Tseng YJ, Wang SY, et al. The metabolome profiling and pathway analysis in metabolic healthy and abnormal obesity. *International journal of obesity* 2015; **39**(8): 1241-8.
78. Oberbach A, Bluher M, Wirth H, et al. Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes. *Journal of proteome research* 2011; **10**(10): 4769-88.
79. Robertson KD. DNA methylation and human disease. *Nat Rev Genet* 2005; **6**(8): 597-610.
80. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002; **16**(1): 6-21.
81. Relton CL, Davey Smith G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS medicine* 2010; **7**(10): e1000356.
82. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet* 2013; **14**(8): 585-94.
83. Richmond RC, Simpkin AJ, Woodward G, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Human molecular genetics* 2015; **24**(8): 2201-17.
84. Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics : official journal of the DNA Methylation Society* 2011; **6**(6): 692-702.
85. Ek WE, Tobi EW, Ahsan M, et al. Tea and coffee consumption in relation to DNA methylation in four European cohorts. *Human molecular genetics* 2017; **26**(16): 3221-31.
86. Ma J, Liu C, Joehanes R, et al. Abstract 12: Whole Blood DNA Methylation Signatures of a Mediterranean-style Dietary Pattern. *Circulation* 2017; **135**(Suppl 1): A12-A.
87. Wang X, Zhu H, Snieder H, et al. Obesity related methylation changes in DNA of peripheral blood leukocytes. *BMC Med* 2010; **8**: 87.
88. Feinberg AP, Irizarry RA, Fradin D, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. *Science translational medicine* 2010; **2**(49): 49ra67.
89. Almen MS, Jacobsson JA, Moschonis G, et al. Genome wide analysis reveals association of a FTO gene variant with epigenetic changes. *Genomics* 2012; **99**(3): 132-7.
90. Xu X, Su S, Barnes VA, et al. A genome-wide methylation study on obesity: differential variability and differential methylation. *Epigenetics : official journal of the DNA Methylation Society* 2013; **8**(5): 522-33.
91. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* 2014; **383**(9933): 1990-8.
92. Agha G, Houseman EA, Kelsey KT, Eaton CB, Buka SL, Loucks EB. Adiposity is associated with DNA methylation profile in adipose tissue. *International journal of epidemiology* 2015; **44**(4): 1277-87.
93. Demerath EW, Guan W, Grove ML, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Human molecular genetics* 2015; **24**(15): 4464-79.
94. Wahl S, Drong A, Lehne B, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* 2017; **541**(7635): 81-6.
95. Mendelson MM, Marioni RE, Joehanes R, et al. Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS medicine* 2017; **14**(1): e1002215.

96. Boyd A, Golding J, Macleod J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *International journal of epidemiology* 2013; **42**(1): 111-27.
97. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International journal of epidemiology* 2013; **42**(1): 97-110.
98. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017; **186**(9): 1026-34.
99. Collins R. What makes UK Biobank special? *Lancet* 2012; **379**(9822): 1173-4.
100. Emmett PM, Jones LR, Northstone K. Dietary patterns in the Avon Longitudinal Study of Parents and Children. *Nutr Rev* 2015; **73 Suppl 3**: 207-30.
101. Northstone K, Smith AD, Cribb VL, Emmett PM. Dietary patterns in UK adolescents obtained from a dual-source FFQ and their associations with socio-economic position, nutrient intake and modes of eating. *Public health nutrition* 2014; **17**(7): 1476-85.
102. Jones LR, Steer CD, Rogers IS, Emmett PM. Influences on child fruit and vegetable intake: sociodemographic, parental and child factors in a longitudinal cohort study. *Public health nutrition* 2010; **13**(7): 1122-30.
103. Smith AD, Emmett PM, Newby PK, Northstone K. Dietary patterns obtained through principal components analysis: the effect of input variable quantification. *The British journal of nutrition* 2013; **109**(10): 1881-91.
104. Paternoster L, Zhurov AI, Toma AM, et al. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *Am J Hum Genet* 2012; **90**(3): 478-85.
105. Granell R, Henderson AJ, Evans DM, et al. Effects of BMI, fat mass, and lean mass on asthma in childhood: a Mendelian randomization study. *PLoS medicine* 2014; **11**(7): e1001669.
106. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015; **526**(7571): 68-74.
107. Relton CL, Gaunt T, McArdle W, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International journal of epidemiology* 2015; **44**(4): 1181-90.
108. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011; **3**(6): 771-84.
109. Richmond RC, Sharp GC, Ward ME, et al. DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework. *Diabetes* 2016; **65**(5): 1231-44.
110. Drenos F, Davey Smith G, Ala-Korpela M, et al. Metabolic Characterization of a Rare Genetic Variation Within APOC3 and Its Lipoprotein Lipase-Independent Effects. *Circ Cardiovasc Genet* 2016; **9**(3): 231-9.
111. Inouye M, Kettunen J, Soininen P, et al. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol* 2010; **6**: 441.
112. Soininen P, Kangas AJ, Wurtz P, et al. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* 2009; **134**(9): 1781-5.
113. Bycroft C, Freeman C, Petkova D, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 2017.
114. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; **48**(10): 1279-83.
115. Mitchell R, Hemani, G, Dudding, T, Paternoster, L. UK Biobank Genetic Data: MRC-IEU Quality Control, Version 1. *University of Bristol* 2017; doi:: 10.5523/bris.3074krb6t2frj29yh2b03x3wxj.
116. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010; **26**(22): 2867-73.

117. Liu B, Young H, Crowe FL, et al. Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. *Public health nutrition* 2011; **14**(11): 1998-2005.
118. Englyst H, Wiggins HS, Cummings JH. Determination of the non-starch polysaccharides in plant foods by gas-liquid chromatography of constituent sugars as alditol acetates. *Analyst* 1982; **107**(1272): 307-18.
119. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 2017; **101**(1): 5-22.
120. Elsworth B, Mitchell, R, Raistrick, CA, Paternoster, L, Hemani, G, Gaunt, TR. MRC IEU UK Biobank GWAS pipeline version 1. *University of Bristol* 2017; doi: 10.5523/bris.3074krb6t2frj29yh2b03x3wxj.
121. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**(3): 559-75.
122. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011; **12**(8): 529-41.
123. Min J, Hemani G, Davey Smith G, Relton CL, Suderman M. Meffil: efficient normalisation and analysis of very large DNA methylation samples. *bioRxiv* 2017.
124. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 2011; **27**(11): 1496-505.
125. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 2014; **30**(10): 1431-9.
126. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *International journal of epidemiology* 2013; **42**(4): 1134-44.
127. Felix JF, Bradfield JP, Monnereau C, et al. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Human molecular genetics* 2016; **25**(2): 389-403.
128. Pierce BL, Burgess S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol* 2013; **178**(7): 1177-84.
129. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011; **478**(7367): 103-9.
130. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013; **37**(7): 658-65.
131. Hemani G, Zheng J, Wade KH, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv* 2016.
132. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology* 2015; **44**(2): 512-25.
133. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. *2014* 2014; **59**(5): 38.
134. Chu AY, Workalemahu T, Paynter NP, et al. Novel locus including FGF21 is associated with dietary macronutrient intake. *Human molecular genetics* 2013; **22**(9): 1895-902.
135. Plomin R, Deary IJ. Genetics and intelligence differences: five special findings. *Mol Psychiatry* 2015; **20**(1): 98-108.
136. Carnell S, Haworth CM, Plomin R, Wardle J. Genetic influence on appetite in children. *International journal of obesity* 2008; **32**(10): 1468-73.
137. Llewellyn CH, Trzaskowski M, van Jaarsveld CHM, Plomin R, Wardle J. Satiety mechanisms in genetic risk of obesity. *JAMA Pediatr* 2014; **168**(4): 338-44.
138. Rietveld CA, Medland SE, Derringer J, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 2013; **340**(6139): 1467-71.

139. Ward ME, McMahon G, St Pourcain B, et al. Genetic variation associated with differential educational attainment in adults has anticipated associations with school performance in children. *PloS one* 2014; **9**(7): e100248.
140. Tanaka T, Ngwa JS, van Rooij FJ, et al. Genome-wide meta-analysis of observational studies shows common genetic variants associated with macronutrient intake. *The American journal of clinical nutrition* 2013; **97**(6): 1395-402.
141. Haghighi A, Melka MG, Bernard M, et al. Opioid receptor mu 1 gene, fat intake and obesity in adolescence. *Mol Psychiatry* 2014; **19**(1): 63-8.
142. Wakai K, Matsuo K, Matsuda F, et al. Genome-wide association study of genetic factors related to confectionery intake: potential roles of the ADIPOQ gene. *Obesity* 2013; **21**(11): 2413-9.
143. Soberg S, Sandholt CH, Jespersen NZ, et al. FGF21 Is a Sugar-Induced Hormone Associated with Sweet Intake and Preference in Humans. *Cell metabolism* 2017; **25**(5): 1045-53 e6.
144. Allen NE, Sudlow C, Peakman T, Collins R, Biobank UK. UK biobank data: come and get it. *Science translational medicine* 2014; **6**(224): 224ed4.
145. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47**(11): 1236-41.
146. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**(3): 291-5.
147. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 2017; **33**(2): 272-9.
148. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010; **42**(11): 937-48.
149. Psychiatric GCBDWG. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 2011; **43**(10): 977-83.
150. van Meurs JB, Pare G, Schwartz SM, et al. Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease. *The American journal of clinical nutrition* 2013; **98**(3): 668-76.
151. Piaggi P, Masindova I, Muller YL, et al. A Genome-Wide Association Study Using a Custom Genotyping Array Identifies Variants in GPR158 Associated With Reduced Energy Expenditure in American Indians. *Diabetes* 2017; **66**(8): 2284-95.
152. Ikram MK, Sim X, Jensen RA, et al. Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. *PLoS genetics* 2010; **6**(10): e1001184.
153. Kharitonenkov A, Shiyanova TL, Koester A, et al. FGF-21 as a novel metabolic regulator. *J Clin Invest* 2005; **115**(6): 1627-35.
154. Kurosu H, Choi M, Ogawa Y, et al. Tissue-specific expression of betaKlotho and fibroblast growth factor (FGF) receptor isoforms determines metabolic activity of FGF19 and FGF21. *J Biol Chem* 2007; **282**(37): 26687-95.
155. Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet* 2014; **46**(6): 543-50.
156. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* 2016; **48**(7): 709-17.
157. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; **466**(7307): 707-13.
158. McGovern DP, Jones MR, Taylor KD, et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Human molecular genetics* 2010; **19**(17): 3468-76.
159. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012; **491**(7422): 119-24.

160. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010; **42**(12): 1118-25.
161. de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* 2017; **49**(2): 256-61.
162. Bustamante M, Standl M, Bassat Q, et al. A genome-wide association meta-analysis of diarrhoeal disease in young children identifies FUT2 locus and provides plausible biological pathways. *Human molecular genetics* 2016; **25**(18): 4127-42.
163. Liang Y, Tang W, Huang T, et al. Genetic variations affecting serum carcinoembryonic antigen levels and status of regional lymph nodes in patients with sporadic colorectal cancer from Southern China. *PloS one* 2014; **9**(6): e97923.
164. Tanaka T, Scheet P, Giusti B, et al. Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations. *Am J Hum Genet* 2009; **84**(4): 477-82.
165. Hazra A, Kraft P, Lazarus R, et al. Genome-wide significant predictors of metabolites in the one-carbon metabolism pathway. *Human molecular genetics* 2009; **18**(23): 4677-87.
166. Chambers JC, Zhang W, Sehmi J, et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet* 2011; **43**(11): 1131-8.
167. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* 2017; **49**(7): 1126-32.
168. Suhre K, Shin SY, Petersen AK, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 2011; **477**(7362): 54-60.
169. Comuzzie AG, Cole SA, Laston SL, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PloS one* 2012; **7**(12): e51954.
170. Li YR, Li J, Zhao SD, et al. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat Med* 2015; **21**(9): 1018-27.
171. Ji SG, Juran BD, Mucha S, et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet* 2017; **49**(2): 269-73.
172. Baurecht H, Hotze M, Brand S, et al. Genome-wide comparative analysis of atopic dermatitis and psoriasis gives insight into opposing genetic mechanisms. *Am J Hum Genet* 2015; **96**(1): 104-20.
173. Tsoi LC, Stuart PE, Tian C, et al. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nat Commun* 2017; **8**: 15382.
174. Weiss FU, Schurmann C, Guenther A, et al. Fucosyltransferase 2 (FUT2) non-secretor status and blood group B are associated with elevated serum lipase activity in asymptomatic subjects, and an increased risk for chronic pancreatitis: a genetic association study. *Gut* 2015; **64**(4): 646-56.
175. He M, Wu C, Xu J, et al. A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9, carcinoembryonic antigen and alpha fetoprotein and their associations with cancer risk. *Gut* 2014; **63**(1): 143-51.
176. Rueedi R, Ledda M, Nicholls AW, et al. Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS genetics* 2014; **10**(2): e1004132.
177. Keene KL, Chen WM, Chen F, et al. Genetic Associations with Plasma B12, B6, and Folate Levels in an Ischemic Stroke Population from the Vitamin Intervention for Stroke Prevention (VISP) Trial. *Front Public Health* 2014; **2**: 112.
178. Hazra A, Kraft P, Selhub J, et al. Common variants of FUT2 are associated with plasma vitamin B12 levels. *Nat Genet* 2008; **40**(10): 1160-2.
179. Lin X, Lu D, Gao Y, et al. Genome-wide association study identifies novel loci associated with serum level of vitamin B12 in Chinese men. *Human molecular genetics* 2012; **21**(11): 2610-7.

180. Nongmaithem SS, Joglekar CV, Krishnaveni GV, et al. GWAS identifies population-specific new regulatory variants in FUT6 associated with plasma B12 concentrations in Indians. *Human molecular genetics* 2017; **26**(13): 2551-64.
181. Wacklin P, Makivuokko H, Alakulppi N, et al. Secretor genotype (FUT2 gene) is strongly associated with the composition of Bifidobacteria in the human intestine. *PloS one* 2011; **6**(5): e20113.
182. Wacklin P, Tuimala J, Nikkila J, et al. Faecal microbiota composition in adults is associated with the FUT2 gene determining the secretor status. *PloS one* 2014; **9**(4): e94863.
183. Rausch P, Rehman A, Kunzel S, et al. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proc Natl Acad Sci U S A* 2011; **108**(47): 19030-5.
184. Ahmeti KB, Ajroud-Driss S, Al-Chalabi A, et al. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. *Neurobiol Aging* 2013; **34**(1): 357 e7-19.
185. Autism Spectrum Disorders Working Group of The Psychiatric Genomics C. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism* 2017; **8**: 21.
186. Davies G, Marioni RE, Liewald DC, et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Mol Psychiatry* 2016; **21**(6): 758-67.
187. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511**(7510): 421-7.
188. Ananthakrishnan AN, Khalili H, Konijeti GG, et al. Long-term intake of dietary fat and risk of ulcerative colitis and Crohn's disease. *Gut* 2014; **63**(5): 776-84.
189. Whitton C, Nicholson SK, Roberts C, et al. National Diet and Nutrition Survey: UK food consumption and nutrient intakes from the first year of the rolling programme and comparisons with previous surveys. *The British journal of nutrition* 2011; **106**(12): 1899-914.
190. Dupuis L, Oudart H, Rene F, Gonzalez de Aguilar JL, Loeffler JP. Evidence for defective energy homeostasis in amyotrophic lateral sclerosis: benefit of a high-energy diet in a transgenic mouse model. *Proc Natl Acad Sci U S A* 2004; **101**(30): 11159-64.
191. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *bioRxiv* 2017.
192. Richardson TG, Zheng J, Davey Smith G, et al. Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk. *Am J Hum Genet* 2017; **101**(4): 590-602.
193. Fortune MD, Guo H, Burren O, et al. Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat Genet* 2015; **47**(7): 839-46.
194. Davey Smith G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes Nutr* 2011; **6**(1): 27-43.
195. Grosso G, Bella F, Godos J, et al. Possible role of diet in cancer: systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk. *Nutr Rev* 2017; **75**(6): 405-19.
196. Mente A, de Koning L, Shannon HS, Anand SS. A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. *Arch Intern Med* 2009; **169**(7): 659-69.
197. Mozaffarian D, Micha R, Wallace S. Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: a systematic review and meta-analysis of randomized controlled trials. *PLoS medicine* 2010; **7**(3): e1000252.
198. O'Neil A, Quirk SE, Housden S, et al. Relationship between diet and mental health in children and adolescents: a systematic review. *Am J Public Health* 2014; **104**(10): e31-42.

199. Paternoster L, Tilling K, Davey Smith G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *PLoS genetics* 2017; **13**(10): e1006944.
200. Cutler GJ, Flood A, Hannan PJ, Slavin JL, Neumark-Sztainer D. Association between major patterns of dietary intake and weight status in adolescents. *The British journal of nutrition* 2012; **108**(2): 349-56.
201. Janssen I, Katzmarzyk PT, Boyce WF, et al. Comparison of overweight and obesity prevalence in school-aged youth from 34 countries and their relationships with physical activity and dietary patterns. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2005; **6**(2): 123-32.
202. Gow ML, Ho M, Burrows TL, et al. Impact of dietary macronutrient distribution on BMI and cardiometabolic outcomes in overweight and obese children and adolescents: a systematic review. *Nutr Rev* 2014; **72**(7): 453-70.
203. Te Morenga L, Mallard S, Mann J. Dietary sugars and body weight: systematic review and meta-analyses of randomised controlled trials and cohort studies. *Bmj* 2012; **346**: e7492.
204. Celis-Morales CA, Lyall DM, Gray SR, et al. Dietary fat and total energy intake modifies the association of genetic profile risk score on obesity: evidence from 48 170 UK Biobank participants. *International journal of obesity* 2017.
205. Reilly JJ, Armstrong J, Dorosty AR, et al. Early life risk factors for obesity in childhood: cohort study. *Bmj* 2005; **330**(7504): 1357.
206. Ambrosini GL, Emmett PM, Northstone K, Howe LD, Tilling K, Jebb SA. Identification of a dietary pattern prospectively associated with increased adiposity during childhood and adolescence. *International journal of obesity* 2012; **36**(10): 1299-305.
207. Johnson L, Mander AP, Jones LR, Emmett PM, Jebb SA. Energy-dense, low-fiber, high-fat dietary pattern is associated with increased fatness in childhood. *The American journal of clinical nutrition* 2008; **87**(4): 846-54.
208. Timpson NJ, Emmett PM, Frayling TM, et al. The fat mass- and obesity-associated locus and dietary intake in children. *The American journal of clinical nutrition* 2008; **88**(4): 971-8.
209. Glynn L, Emmett P, Rogers I, Team AS. Food and nutrient intakes of a population sample of 7-year-old children in the south-west of England in 1999/2000 - what difference does gender make? *J Hum Nutr Diet* 2005; **18**(1): 7-19; quiz 21-3.
210. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG, Consortium E-I. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European journal of epidemiology* 2015; **30**(7): 543-52.
211. Relton CL, Davey Smith G. Mendelian randomization: applications and limitations in epigenetic studies. *Epigenomics* 2015; **7**(8): 1239-43.
212. Richmond RC, Davey Smith G, Ness AR, den Hoed M, McMahon G, Timpson NJ. Assessing causality in the association between child adiposity and physical activity levels: a Mendelian randomization analysis. *PLoS medicine* 2014; **11**(3): e1001618.
213. Wardle J, Carnell S, Haworth CM, Farooqi IS, O'Rahilly S, Plomin R. Obesity associated genetic variation in FTO is associated with diminished satiety. *The Journal of clinical endocrinology and metabolism* 2008; **93**(9): 3640-3.
214. Wurtz P, Wang Q, Niironen M, et al. Metabolic signatures of birthweight in 18 288 adolescents and adults. *International journal of epidemiology* 2016; **45**(5): 1539-50.
215. Perrone RD, Madias NE, Levey AS. Serum creatinine as an index of renal function: new insights into old concepts. *Clin Chem* 1992; **38**(10): 1933-53.
216. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med* 1999; **130**(6): 461-70.
217. Coresh J, Selvin E, Stevens LA, et al. Prevalence of chronic kidney disease in the United States. *JAMA : the journal of the American Medical Association* 2007; **298**(17): 2038-47.

218. Hivert MF, Perng W, Watkins SM, et al. Metabolomics in the developmental origins of obesity and its cardiometabolic consequences. *J Dev Orig Health Dis* 2015; **6**(2): 65-78.
219. Wittenbecher C, Muhlenbruch K, Kroger J, et al. Amino acids, lipid metabolites, and ferritin as potential mediators linking red meat consumption to type 2 diabetes. *The American journal of clinical nutrition* 2015; **101**(6): 1241-50.
220. Gibbons H, Brennan L. Metabolomics as a tool in the identification of dietary biomarkers. *The Proceedings of the Nutrition Society* 2017; **76**(1): 42-53.
221. Brennan L. NMR-based metabolomics: from sample preparation to applications in nutrition research. *Prog Nucl Magn Reson Spectrosc* 2014; **83**: 42-9.
222. MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding and suppression effect. *Prev Sci* 2000; **1**(4): 173-81.
223. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annu Rev Psychol* 2007; **58**: 593-614.
224. Kettunen J, Demirkan A, Wurtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* 2016; **7**: 11122.
225. Meyer BJ, Mann NJ, Lewis JL, Milligan GC, Sinclair AJ, Howe PR. Dietary intakes and food sources of omega-6 and omega-3 polyunsaturated fatty acids. *Lipids* 2003; **38**(4): 391-8.
226. Dhiman TR, Satter LD, Pariza MW, Galli MP, Albright K, Tolosa MX. Conjugated linoleic acid (CLA) content of milk from cows offered diets rich in linoleic and linolenic acid. *J Dairy Sci* 2000; **83**(5): 1016-27.
227. Zeilinger S, Kuhnel B, Klopp N, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PloS one* 2013; **8**(5): e63812.
228. Elliott HR, Tillin T, McArdle WL, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics* 2014; **6**(1): 4.
229. Zhang FF, Morabia A, Carroll J, et al. Dietary patterns are associated with levels of global genomic DNA methylation in a cancer-free population. *The Journal of nutrition* 2011; **141**(6): 1165-71.
230. Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. *Ann Epidemiol* 2007; **17**(7): 511-3.
231. Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* 2016; **17**: 61.
232. Rzehak P, Covic M, Saffery R, et al. DNA-Methylation and Body Composition in Preschool Children: Epigenome-Wide-Analysis in the European Childhood Obesity Project (CHOP)-Study. *Sci Rep* 2017; **7**(1): 14349.
233. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 2016; **167**(5): 1415-29 e19.
234. Gieger C, Radhakrishnan A, Cvejic A, et al. New gene functions in megakaryopoiesis and platelet formation. *Nature* 2011; **480**(7376): 201-8.
235. Vizioli L, Muscari S, Muscari A. The relationship of mean platelet volume with the risk and prognosis of cardiovascular diseases. *Int J Clin Pract* 2009; **63**(10): 1509-15.

Appendix A

Appendix A.1 – Metabolite transformations

List of metabolites that were log-transformed in the ALSPAC children and adolescents.

Category	Name etc.
Chylomicrons and extremely large VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Very large VLDL	Particle
	Lipid
	Phospholipids
	Cholesterol
	Cholesterol esters
	Free cholesterol
	Triglycerides
Large VLDL	Free cholesterol
Medium LDL	Triglycerides
Small LDL	Triglycerides
Glycerides & phospholipids	Diacylglycerol
	Ratio of diacylglycerol to triglycerides
Fatty acids & saturation	Conjugated linoleic acid (CLA)
	CLA to total FAs ratio
Amino acids	Histidine
Ketone bodies	Acetoacetate
	3-hydroxybutyrate

Appendix A.2 – MR-Egger results table

Results from MR-Egger analyses performed to investigate the validity of using the GIANT BMI score to assess the causal effect of BMI on the metabolome in the ALSPAC children at age 7 years (**5.2.4**). Effect estimates are the 1-SD increase in metabolite concentration per 1kg/m^2 increase in BMI.

Category	Metabolite	Beta	95% CI	p-value
Chylomicrons and extremely large VLDL	**Particle**	-6.46×10^{-4}	-1.90×10^{-3}, 6.07×10^{-4}	0.312
	Lipid	-6.31×10^{-4}	-1.88×10^{-3}, 6.16×10^{-4}	0.321
	Phospholipids	-6.49×10^{-4}	-1.88×10^{-3}, 5.85×10^{-4}	0.302
	Cholesterol	-7.37×10^{-4}	-2.02×10^{-3}, 5.45×10^{-4}	0.260
	Cholesterol esters	-7.48×10^{-4}	-2.06×10^{-3}, 5.66×10^{-4}	0.265
	Free cholesterol	-6.94×10^{-4}	-1.93×10^{-3}, 5.45×10^{-4}	0.272
	Triglycerides	-6.10×10^{-4}	-1.85×10^{-3}, 6.29×10^{-4}	0.335
Very large VLDL	**Particle**	-6.98×10^{-4}	-1.92×10^{-3}, 5.27×10^{-4}	0.264
	Lipid	-6.74×10^{-4}	-1.91×10^{-3}, 5.58×10^{-4}	0.284
	Phospholipids	-7.26×10^{-4}	-1.95×10^{-3}, 5.00×10^{-4}	0.246
	Cholesterol	-7.11×10^{-4}	-1.99×10^{-3}, 5.64×10^{-4}	0.274
	Cholesterol esters	-7.13×10^{-4}	-2.00×10^{-3}, 5.75×10^{-4}	0.278
	Free cholesterol	-7.10×10^{-4}	-1.97×10^{-3}, 5.50×10^{-4}	0.269
	Triglycerides	-6.70×10^{-4}	-1.89×10^{-3}, 5.50×10^{-4}	0.282
Large VLDL	**Particle**	-7.36×10^{-4}	-1.96×10^{-3}, 4.91×10^{-4}	0.240
	Lipid	-7.52×10^{-4}	-1.98×10^{-3}, 4.81×10^{-4}	0.232
	Phospholipids	-7.72×10^{-4}	-2.00×10^{-3}, 4.60×10^{-4}	0.219
	Cholesterol	-8.01×10^{-4}	-2.06×10^{-3}, 4.58×10^{-4}	0.212
	Cholesterol esters	-8.19×10^{-4}	-2.11×10^{-3}, 4.71×10^{-4}	0.213
	Free cholesterol	-7.62×10^{-4}	-1.99×10^{-3}, 4.69×10^{-4}	0.225
	Triglycerides	-7.24×10^{-4}	-1.95×10^{-3}, 5.00×10^{-4}	0.246
Medium VLDL	**Particle**	-7.87×10^{-4}	-2.05×10^{-3}, 4.72×10^{-4}	0.221
	Lipid	-8.01×10^{-4}	-2.07×10^{-3}, 4.72×10^{-4}	0.218
	Phospholipids	-8.33×10^{-4}	-2.11×10^{-3}, 4.42×10^{-4}	0.200
	Cholesterol	-8.13×10^{-4}	-2.14×10^{-3}, 5.14×10^{-4}	0.230
	Cholesterol esters	-7.12×10^{-4}	-2.08×10^{-3}, 6.52×10^{-4}	0.307
	Free cholesterol	-8.53×10^{-4}	-2.12×10^{-3}, 4.10×10^{-4}	0.186
	Triglycerides	-7.66×10^{-4}	-2.02×10^{-3}, 4.85×10^{-4}	0.230
Small VLDL	**Particle**	-8.24×10^{-4}	-2.15×10^{-3}, 5.05×10^{-4}	0.224
	Lipid	-7.92×10^{-4}	-2.18×10^{-3}, 6.00×10^{-4}	0.265
	Phospholipids	-8.44×10^{-4}	-2.19×10^{-3}, 5.05×10^{-4}	0.220
	Cholesterol	-5.45×10^{-4}	-2.13×10^{-3}, 1.03×10^{-3}	0.499
	Cholesterol esters	-3.45×10^{-4}	-1.99×10^{-3}, 1.30×10^{-3}	0.681
	Free cholesterol	-8.34×10^{-4}	-2.23×10^{-3}, 5.64×10^{-4}	0.242
	Triglycerides	-8.45×10^{-4}	-2.11×10^{-3}, 4.23×10^{-4}	0.192
Very small VLDL	**Particle**	-5.54×10^{-4}	-2.10×10^{-3}, 9.94×10^{-4}	0.483
	Lipid	-2.70×10^{-4}	-1.89×10^{-3}, 1.35×10^{-3}	0.744
	Phospholipids	-4.04×10^{-4}	-2.01×10^{-3}, 1.20×10^{-3}	0.621
	Cholesterol	2.17×10^{-4}	-1.30×10^{-3}, 1.74×10^{-3}	0.780
	Cholesterol esters	2.92×10^{-4}	-1.29×10^{-3}, 1.87×10^{-3}	0.717
	Triglycerides	-9.76×10^{-4}	-2.27×10^{-3}, 3.17×10^{-4}	0.139
IDL	**Cholesterol**	-1.99×10^{-4}	-1.87×10^{-3}, 1.47×10^{-3}	0.815
	Cholesterol esters	-2.23×10^{-4}	-1.91×10^{-3}, 1.47×10^{-3}	0.796
Medium LDL	**Phospholipids**	-7.83×10^{-4}	-2.39×10^{-3}, 8.27×10^{-4}	0.341
Small LDL	**Phospholipids**	-7.54×10^{-4}	-2.29×10^{-3}, 7.81×10^{-4}	0.335
Very large HDL	**Particle**	-3.62×10^{-4}	-1.54×10^{-3}, 8.19×10^{-4}	0.548
	Lipid	-3.19×10^{-4}	-1.48×10^{-3}, 8.42×10^{-4}	0.590
	Phospholipids	-4.02×10^{-4}	-1.59×10^{-3}, 7.90×10^{-4}	0.509
	Cholesterol	-1.66×10^{-4}	-1.29×10^{-3}, 9.61×10^{-4}	0.773
	Cholesterol esters	-9.58×10^{-5}	-1.22×10^{-3}, 1.03×10^{-3}	0.867
	Free cholesterol	-3.39×10^{-4}	-1.48×10^{-3}, 8.05×10^{-4}	0.561
Large HDL	**Particle**	-3.95×10^{-4}	-1.70×10^{-3}, 9.08×10^{-4}	0.552
	Lipid	-4.45×10^{-4}	-1.74×10^{-3}, 8.49×10^{-4}	0.500
	Phospholipids	-5.19×10^{-4}	-1.79×10^{-3}, 7.55×10^{-4}	0.425
	Cholesterol	-3.67×10^{-4}	-1.68×10^{-3}, 9.44×10^{-4}	0.583
	Cholesterol esters	-3.73×10^{-4}	-1.69×10^{-3}, 9.46×10^{-4}	0.580
	Free cholesterol	-3.57×10^{-4}	-1.64×10^{-3}, 9.24×10^{-4}	0.585

Medium HDL	**Cholesterol**	-8.68×10^{-4}	-2.13×10^{-3}, 3.97×10^{-4}	0.179
	Cholesterol esters	-8.06×10^{-4}	-2.10×10^{-3}, 4.89×10^{-4}	0.222
	Free cholesterol	-1.07×10^{-3}	-2.23×10^{-3}, 9.60×10^{-5}	0.072
Small HDL	**Particle**	-9.43×10^{-4}	-2.11×10^{-3}, 2.25×10^{-4}	0.114
	Lipid	-1.04×10^{-3}	-2.21×10^{-3}, 1.21×10^{-4}	0.079
	Phospholipids	-7.47×10^{-4}	-2.01×10^{-3}, 5.12×10^{-4}	0.245
	Cholesterol	-7.61×10^{-4}	-2.00×10^{-3}, 4.81×10^{-4}	0.230
	Cholesterol esters	-6.73×10^{-4}	-1.97×10^{-3}, 6.19×10^{-4}	0.307
	Triglycerides	-7.40×10^{-4}	-1.91×10^{-3}, 4.34×10^{-4}	0.217
Lipoprotein particle sizes	**VLDL particle size**	-4.12×10^{-4}	-1.64×10^{-3}, 8.19×10^{-4}	0.512
	HDL particle size	-2.06×10^{-4}	-1.45×10^{-3}, 1.03×10^{-3}	0.744
Cholesterol	**VLDL cholesterol**	-6.13×10^{-4}	-2.10×10^{-3}, 8.73×10^{-4}	0.419
	Remnant cholesterol	-5.08×10^{-4}	-2.16×10^{-3}, 1.14×10^{-3}	0.547
	HDL cholesterol	-5.53×10^{-4}	-1.80×10^{-3}, 6.97×10^{-4}	0.386
	HDL2 cholesterol	-3.86×10^{-4}	-1.66×10^{-3}, 8.90×10^{-4}	0.553
	HDL3 cholesterol	-8.26×10^{-4}	-2.03×10^{-3}, 3.79×10^{-4}	0.179
Glycerides & phospholipids	**Triglycerides**	-8.60×10^{-4}	-2.11×10^{-3}, 3.85×10^{-4}	0.176
	VLDL triglycerides	-7.81×10^{-4}	-2.03×10^{-3}, 4.66×10^{-4}	0.220
	HDL triglycerides	-9.03×10^{-4}	-2.09×10^{-3}, 2.82×10^{-4}	0.135
	Diacylglycerol	-1.18×10^{-3}	-2.49×10^{-3}, 1.38×10^{-4}	0.079
	Ratio of diacylglycerol to triglycerides	-8.72×10^{-4}	-2.14×10^{-3}, 4.01×10^{-4}	0.179
Apolipoproteins	**ApoA-I**	-9.28×10^{-4}	-2.13×10^{-3}, 2.74×10^{-4}	0.130
	ApoB	-7.15×10^{-4}	-2.29×10^{-3}, 8.55×10^{-4}	0.372
	ApoB/ApoA-I	-2.80×10^{-4}	-1.84×10^{-3}, 1.28×10^{-3}	0.725
Fatty acids & saturation	**Total fatty acids (FA)**	-1.26×10^{-3}	-2.59×10^{-3}, 7.49×10^{-5}	0.064
	Estimated fatty acid chain length	-3.38×10^{-5}	-1.10×10^{-3}, 1.04×10^{-3}	0.951
	Docosahexaenoic acids (DHA)	-1.17×10^{-3}	-2.43×10^{-3}, 8.84×10^{-5}	0.068
	Conjugated linoleic acid (CLA)	-6.53×10^{-4}	-1.95×10^{-3}, 6.43×10^{-4}	0.323
	Omega-3 fatty acids	-1.24×10^{-3}	-2.51×10^{-3}, 2.85×10^{-5}	0.055
	Omega-6 fatty acids	-1.06×10^{-3}	-2.48×10^{-3}, 3.59×10^{-4}	0.143
	PUFA	-1.13×10^{-3}	-2.54×10^{-3}, 2.74×10^{-4}	0.115
	MUFA	-1.25×10^{-3}	-2.51×10^{-3}, 1.22×10^{-5}	0.052
	Saturated fatty acids (SFA)	-1.03×10^{-3}	-2.30×10^{-3}, 2.42×10^{-4}	0.112
	LA to total FAs ratio	4.77×10^{-4}	-7.95×10^{-4}, 1.75×10^{-3}	0.463
	CLA to total FAs ratio	-6.02×10^{-4}	-1.88×10^{-3}, 6.72×10^{-4}	0.355
	Omega-6 to total FAs ratio	4.22×10^{-4}	-7.67×10^{-4}, 1.61×10^{-3}	0.486
	PUFAs to total FAs ratio	2.88×10^{-4}	-8.98×10^{-4}, 1.47×10^{-3}	0.634
	MUFAs to total FAs ratio	-5.99×10^{-4}	-1.79×10^{-3}, 5.91×10^{-4}	0.324
Glycolysis related metabolites	**Glucose**	-1.30×10^{-4}	-1.37×10^{-3}, 1.11×10^{-3}	0.837
	Lactate	-5.01×10^{-4}	-1.66×10^{-3}, 6.57×10^{-4}	0.397
	Citrate	2.01×10^{-4}	-8.43×10^{-4}, 1.24×10^{-3}	0.706
Amino acids	**Glutamine**	-9.89×10^{-4}	-2.08×10^{-3}, 1.03×10^{-4}	0.076
	Histidine	-6.48×10^{-4}	-1.70×10^{-3}, 4.00×10^{-4}	0.226
	Isoleucine	-3.02×10^{-4}	-1.34×10^{-3}, 7.41×10^{-4}	0.570
	Leucine	7.58×10^{-5}	-9.46×10^{-4}, 1.10×10^{-3}	0.884
	Valine	2.66×10^{-4}	-8.01×10^{-4}, 1.33×10^{-3}	0.625
	Phenylalanine	-2.89×10^{-4}	-1.39×10^{-3}, 8.11×10^{-4}	0.606
	Tyrosine	1.10×10^{-3}	4.27×10^{-5}, 2.15×10^{-3}	0.041
Ketone bodies	**3-hydroxybutyrate**	-9.88×10^{-4}	-2.07×10^{-3}, 9.51×10^{-5}	0.074
Fluid balance	**Creatinine**	-1.57×10^{-3}	-2.68×10^{-3}, -4.61×10^{-4}	0.006
Inflammation	**Glycoprotein acetyls**	-9.56×10^{-4}	-2.25×10^{-3}, 3.32×10^{-4}	0.146

Appendix A.3 – Cross-sectional associations between diet PCs and metabolites

Results from cross-sectional analyses of the relationships between the diet PCs and metabolites in the ALSPAC children at age 7 years (**6.2.1**). Effect estimates are the 1-SD increase in metabolite concentration per unit increase in diet PC.

Category	Metabolite	Health aware PC			Packed lunch PC		
		Beta	95% CI	p-value	Beta	95% CI	p-value
Chylomicrons and extremely large VLDL	**Particle**	-0.002	-0.02, 0.015	0.799	-0.025	-0.046, -0.005	0.013
	Lipid	-0.002	-0.02, 0.015	0.795	-0.025	-0.045, -0.004	0.017
	Phospholipids	-0.003	-0.02, 0.015	0.759	-0.026	-0.046, -0.005	0.013
	Cholesterol	-0.006	-0.024, 0.011	0.489	-0.022	-0.042, -0.002	0.034
	Cholesterol esters	-0.010	-0.027, 0.008	0.282	-0.017	-0.037, 0.003	0.097
	Free cholesterol	-0.003	-0.02, 0.015	0.767	-0.025	-0.046, -0.005	0.014
	Triglycerides	-0.002	-0.019, 0.016	0.865	-0.025	-0.045, -0.005	0.016
Very large VLDL	**Particle**	-0.001	-0.019, 0.016	0.872	-0.020	-0.041, 0	0.049
	Lipid	-0.002	-0.019, 0.016	0.857	-0.020	-0.041, 0	0.050
	Phospholipids	-0.002	-0.02, 0.015	0.787	-0.023	-0.043, -0.003	0.027
	Cholesterol	-0.003	-0.021, 0.015	0.739	-0.022	-0.042, -0.002	0.033
	Cholesterol esters	-0.003	-0.02, 0.015	0.758	-0.019	-0.039, 0.001	0.064
	Free cholesterol	-0.003	-0.021, 0.014	0.718	-0.025	-0.045, -0.004	0.017
	Triglycerides	-0.001	-0.019, 0.017	0.890	-0.019	-0.039, 0.002	0.070
Large VLDL	**Particle**	-0.004	-0.021, 0.014	0.696	-0.014	-0.034, 0.006	0.175
	Lipid	-0.003	-0.021, 0.014	0.699	-0.014	-0.034, 0.006	0.166
	Phospholipids	-0.003	-0.021, 0.014	0.713	-0.016	-0.036, 0.004	0.122
	Cholesterol	-0.003	-0.021, 0.014	0.699	-0.016	-0.036, 0.005	0.130
	Cholesterol esters	-0.004	-0.022, 0.013	0.640	-0.013	-0.033, 0.007	0.205
	Free cholesterol	-0.003	-0.02, 0.015	0.766	-0.018	-0.038, 0.003	0.086
	Triglycerides	-0.004	-0.021, 0.014	0.697	-0.013	-0.033, 0.007	0.198
Medium VLDL	**Particle**	-0.003	-0.021, 0.015	0.732	-0.011	-0.032, 0.009	0.267
	Lipid	-0.003	-0.02, 0.015	0.749	-0.011	-0.031, 0.009	0.282
	Phospholipids	-0.003	-0.02, 0.015	0.755	-0.013	-0.033, 0.007	0.205
	Cholesterol	0.000	-0.018, 0.017	0.990	-0.012	-0.032, 0.008	0.251
	Cholesterol esters	0.002	-0.015, 0.02	0.797	-0.009	-0.029, 0.011	0.374
	Free cholesterol	-0.003	-0.02, 0.015	0.768	-0.014	-0.034, 0.006	0.182
	Triglycerides	-0.004	-0.022, 0.014	0.653	-0.010	-0.03, 0.01	0.343
Small VLDL	**Particle**	-0.009	-0.027, 0.008	0.306	-0.011	-0.031, 0.009	0.289
	Lipid	-0.012	-0.029, 0.005	0.179	-0.011	-0.031, 0.009	0.278
	Phospholipids	-0.020	-0.037, -0.002	0.026	-0.005	-0.025, 0.015	0.601
	Cholesterol	-0.014	-0.031, 0.004	0.125	-0.017	-0.037, 0.003	0.099
	Cholesterol esters	-0.013	-0.031, 0.004	0.138	-0.019	-0.039, 0.001	0.066
	Free cholesterol	-0.013	-0.03, 0.005	0.159	-0.011	-0.031, 0.009	0.279
	Triglycerides	-0.006	-0.024, 0.012	0.504	-0.008	-0.028, 0.012	0.454
Very small VLDL	**Particle**	0.001	-0.016, 0.019	0.869	-0.018	-0.038, 0.002	0.077
	Lipid	-0.001	-0.018, 0.017	0.954	-0.020	-0.04, 0	0.051
	Phospholipids	0.007	-0.01, 0.025	0.415	-0.029	-0.049, -0.009	0.005
	Cholesterol	0.000	-0.017, 0.018	0.966	-0.011	-0.031, 0.01	0.303
	Cholesterol esters	-0.003	-0.02, 0.015	0.772	-0.015	-0.035, 0.006	0.158
	Free cholesterol	0.006	-0.012, 0.024	0.489	-0.002	-0.022, 0.019	0.860
	Triglycerides	-0.012	-0.029, 0.006	0.193	-0.014	-0.034, 0.006	0.168

IDL	**Particle**	0.015	-0.003, 0.032	0.109	-0.043	-0.064, -0.023	3.21×10^{-5}
	Lipid	0.012	-0.006, 0.029	0.202	-0.037	-0.058, -0.017	3.38×10^{-4}
	Phospholipids	0.013	-0.005, 0.031	0.151	-0.041	-0.061, -0.02	8.91×10^{-5}
	Cholesterol	0.014	-0.004, 0.032	0.123	-0.034	-0.054, -0.014	0.001
	Cholesterol esters	0.012	-0.006, 0.03	0.180	-0.032	-0.052, -0.011	0.002
	Free cholesterol	0.017	-0.001, 0.035	0.057	-0.037	-0.058, -0.017	4.08×10^{-4}
	Triglycerides	-0.012	-0.03, 0.006	0.186	-0.027	-0.048, -0.007	0.009
Large LDL	**Particle**	0.006	-0.012, 0.024	0.505	-0.042	-0.063, -0.022	5.47×10^{-5}
	Lipid	0.006	-0.011, 0.024	0.482	-0.040	-0.061, -0.02	1.18×10^{-4}
	Phospholipids	0.004	-0.014, 0.022	0.660	-0.038	-0.059, -0.018	2.20×10^{-4}
	Cholesterol	0.010	-0.008, 0.027	0.294	-0.039	-0.06, -0.019	1.69×10^{-4}
	Cholesterol esters	0.009	-0.008, 0.027	0.304	-0.040	-0.06, -0.019	1.41×10^{-4}
	Free cholesterol	0.010	-0.008, 0.028	0.268	-0.038	-0.058, -0.017	3.36×10^{-4}
	Triglycerides	-0.014	-0.032, 0.004	0.135	-0.036	-0.057, -0.015	0.001
Medium LDL	**Particle**	0.003	-0.015, 0.021	0.739	-0.042	-0.063, -0.022	4.87×10^{-5}
	Lipid	0.004	-0.014, 0.022	0.657	-0.041	-0.061, -0.02	8.72×10^{-5}
	Phospholipids	-0.010	-0.028, 0.008	0.265	-0.034	-0.055, -0.014	0.001
	Cholesterol	0.010	-0.008, 0.028	0.257	-0.041	-0.061, -0.02	9.09×10^{-5}
	Cholesterol esters	0.013	-0.005, 0.03	0.168	-0.041	-0.062, -0.021	7.18×10^{-5}
	Free cholesterol	0.000	-0.018, 0.018	0.961	-0.036	-0.057, -0.016	0.001
	Triglycerides	-0.016	-0.034, 0.002	0.085	-0.036	-0.056, -0.015	0.001
Small LDL	**Particle**	-0.001	-0.019, 0.017	0.914	-0.043	-0.063, -0.022	4.21×10^{-5}
	Lipid	0.003	-0.015, 0.021	0.736	-0.041	-0.061, -0.02	9.91×10^{-5}
	Phospholipids	-0.007	-0.025, 0.011	0.464	-0.038	-0.058, -0.017	3.09×10^{-4}
	Cholesterol	0.008	-0.01, 0.026	0.361	-0.040	-0.06, -0.019	1.46×10^{-4}
	Cholesterol esters	0.011	-0.007, 0.029	0.221	-0.041	-0.061, -0.02	9.30×10^{-5}
	Free cholesterol	-0.006	-0.025, 0.012	0.480	-0.027	-0.048, -0.006	0.010
	Triglycerides	-0.016	-0.034, 0.002	0.075	-0.033	-0.054, -0.013	0.001
Very large HDL	**Particle**	0.022	0.003, 0.04	0.020	-0.032	-0.053, -0.011	0.002
	Lipid	0.020	0.002, 0.038	0.031	-0.035	-0.056, -0.014	0.001
	Phospholipids	0.025	0.007, 0.044	0.007	-0.027	-0.048, -0.006	0.011
	Cholesterol	0.011	-0.007, 0.029	0.233	-0.039	-0.06, -0.019	2.23×10^{-4}
	Cholesterol esters	0.007	-0.011, 0.025	0.463	-0.041	-0.062, -0.02	1.49×10^{-4}
	Free cholesterol	0.021	0.003, 0.04	0.022	-0.035	-0.056, -0.014	0.001
	Triglycerides	0.032	0.014, 0.05	0.001	-0.040	-0.061, -0.019	1.69×10^{-4}
Large HDL	**Particle**	0.016	-0.002, 0.034	0.081	-0.013	-0.034, 0.008	0.213
	Lipid	0.017	-0.001, 0.035	0.063	-0.011	-0.032, 0.009	0.284
	Phospholipids	0.015	-0.003, 0.033	0.098	-0.015	-0.036, 0.006	0.153
	Cholesterol	0.017	-0.001, 0.035	0.069	-0.008	-0.028, 0.013	0.473
	Cholesterol esters	0.016	-0.002, 0.034	0.085	-0.007	-0.028, 0.014	0.505
	Free cholesterol	0.020	0.002, 0.038	0.032	-0.010	-0.03, 0.011	0.367
	Triglycerides	0.054	0.036, 0.073	5.47×10^{-9}	-0.024	-0.044, -0.003	0.027
Medium HDL	**Particle**	-0.039	-0.057, -0.021	2.16×10^{-5}	0.015	-0.006, 0.036	0.153
	Lipid	-0.038	-0.056, -0.02	4.55×10^{-5}	0.016	-0.005, 0.036	0.139
	Phospholipids	-0.044	-0.062, -0.026	2.24×10^{-6}	0.011	-0.01, 0.032	0.298
	Cholesterol	-0.026	-0.044, -0.008	0.005	0.021	0, 0.042	0.048
	Cholesterol esters	-0.028	-0.046, -0.01	0.003	0.023	0.002, 0.044	0.030
	Free cholesterol	-0.017	-0.036, 0.001	0.063	0.012	-0.009, 0.033	0.256
	Triglycerides	-0.015	-0.033, 0.003	0.094	-0.005	-0.025, 0.016	0.644
Small HDL	**Particle**	-0.067	-0.085, -0.049	6.10×10^{-13}	0.015	-0.006, 0.036	0.162
	Lipid	-0.070	-0.088, -0.052	5.57×10^{-14}	0.016	-0.005, 0.036	0.141
	Phospholipids	-0.052	-0.071, -0.034	1.68×10^{-8}	0.014	-0.007, 0.035	0.190
	Cholesterol	-0.053	-0.072, -0.035	7.89×10^{-9}	0.014	-0.007, 0.035	0.185
	Cholesterol esters	-0.047	-0.065, -0.029	4.00×10^{-7}	0.012	-0.009, 0.032	0.278
	Free cholesterol	-0.052	-0.07, -0.034	1.40×10^{-8}	0.017	-0.003, 0.038	0.102
	Triglycerides	-0.023	-0.041, -0.005	0.012	-0.017	-0.037, 0.004	0.107
Lipoprotein particle sizes	**VLDL particle size**	-0.003	-0.021, 0.015	0.711	-0.010	-0.031, 0.01	0.332
	LDL particle size	0.035	0.017, 0.054	1.48×10^{-4}	0.017	-0.004, 0.038	0.112
	HDL particle size	0.034	0.016, 0.052	2.22×10^{-4}	-0.023	-0.043, -0.002	0.033

Cholesterol	**Total cholesterol**	0.005	-0.013, 0.023	0.581	-0.036	-0.057, -0.016	0.001
	VLDL cholesterol	-0.005	-0.022, 0.012	0.584	-0.017	-0.037, 0.003	0.089
	Remnant cholesterol	0.004	-0.014, 0.021	0.671	-0.028	-0.048, -0.008	0.006
	LDL cholesterol	0.010	-0.008, 0.027	0.293	-0.040	-0.06, -0.02	1.29×10^{-4}
	HDL cholesterol	-0.005	-0.023, 0.013	0.578	-0.009	-0.03, 0.012	0.405
	HDL2 cholesterol	-0.008	-0.026, 0.01	0.365	-0.006	-0.026, 0.015	0.592
	HDL3 cholesterol	0.001	-0.017, 0.02	0.901	-0.014	-0.035, 0.007	0.187
	Esterified cholesterol	0.009	-0.009, 0.027	0.316	-0.035	-0.056, -0.015	0.001
	Free cholesterol	-0.005	-0.023, 0.013	0.572	-0.036	-0.057, -0.016	0.001
Glycerides & phospholipids	**Triglycerides**	-0.006	-0.024, 0.011	0.488	-0.018	-0.038, 0.002	0.080
	VLDL triglycerides	-0.004	-0.022, 0.013	0.625	-0.013	-0.033, 0.008	0.222
	LDL triglycerides	-0.015	-0.033, 0.003	0.102	-0.036	-0.057, -0.015	0.001
	HDL triglycerides	0.002	-0.016, 0.02	0.849	-0.023	-0.043, -0.002	0.029
	Diacylglycerol	0.014	-0.005, 0.032	0.144	-0.011	-0.032, 0.01	0.310
	Ratio of diacylglycerol to triglycerides	0.010	-0.008, 0.029	0.262	-0.002	-0.023, 0.019	0.827
	Phosphoglycerides	0.019	0.001, 0.038	0.038	-0.022	-0.043, -0.001	0.039
	Ratio of triglycerides to phophoglycerides	0.006	-0.012, 0.024	0.501	-0.009	-0.029, 0.012	0.413
	Phosphatidylcholine and other cholines	0.025	0.007, 0.044	0.007	-0.025	-0.046, -0.004	0.022
	Total cholines	0.019	0.001, 0.038	0.039	-0.024	-0.045, -0.003	0.023
Apolipoproteins	**ApoA-I**	-0.005	-0.024, 0.013	0.581	-0.018	-0.039, 0.003	0.100
	ApoB	0.003	-0.014, 0.02	0.737	-0.031	-0.051, -0.011	0.002
	ApoB/ApoA-I	0.005	-0.013, 0.022	0.589	-0.023	-0.043, -0.003	0.024
Fatty acids & saturation	**Total fatty acids (FA)**	0.014	-0.004, 0.032	0.126	-0.024	-0.045, -0.004	0.019
	Estimated fatty acid chain length	-0.008	-0.026, 0.011	0.411	-0.002	-0.023, 0.018	0.817
	Estimated degree of unsaturation	0.083	0.065, 0.101	7.59×10^{-19}	0.001	-0.02, 0.022	0.912
	Docosahexaenoic acids (DHA)	0.127	0.109, 0.145	2.86×10^{-43}	-0.110	-0.131, -0.09	1.70×10^{-25}
	Linoleic acid (LA)	0.038	0.02, 0.056	4.20×10^{-5}	0.023	0.003, 0.044	0.027
	Conjugated linoleic acid (CLA)	0.045	0.027, 0.063	1.21×10^{-6}	-0.069	-0.09, -0.048	7.61×10^{-11}
	Omega-3 fatty acids	0.099	0.081, 0.117	5.19×10^{-27}	-0.099	-0.119, -0.078	6.50×10^{-21}
	Omega-6 fatty acids	0.047	0.029, 0.065	2.74×10^{-7}	0.003	-0.018, 0.024	0.780
	PUFA	0.056	0.038, 0.074	1.06×10^{-9}	-0.010	-0.03, 0.011	0.346
	MUFA	-0.042	-0.059, -0.024	3.23×10^{-6}	-0.020	-0.04, 0	0.050
	Saturated fatty acids (SFA)	0.024	0.006, 0.042	0.010	-0.033	-0.054, -0.013	0.002
	DHA to total FAs ratio	0.140	0.122, 0.158	4.62×10^{-51}	-0.115	-0.135, -0.094	8.78×10^{-27}
	LA to total FAs ratio	0.042	0.023, 0.06	7.75×10^{-6}	0.086	0.065, 0.106	6.33×10^{-16}
	CLA to total FAs ratio	0.043	0.025, 0.062	3.49×10^{-6}	-0.074	-0.095, -0.053	4.70×10^{-12}
	Omega-3 to total FAs ratio	0.114	0.096, 0.132	2.70×10^{-34}	-0.105	-0.126, -0.084	1.44×10^{-22}
	Omega-6 to total FAs ratio	0.060	0.042, 0.078	1.00×10^{-10}	0.053	0.033, 0.074	4.28×10^{-7}
	PUFAs to total FAs ratio	0.078	0.06, 0.096	2.03×10^{-17}	0.031	0.01, 0.052	0.003
	MUFAs to total FAs ratio	-0.105	-0.123, -0.088	3.49×10^{-31}	-0.009	-0.03, 0.011	0.388
	SFAs to total FAs ratio	0.039	0.02, 0.057	3.14×10^{-5}	-0.034	-0.055, -0.013	0.001
Glycolysis related metabolites	**Glucose**	-0.011	-0.029, 0.008	0.258	0.013	-0.008, 0.034	0.213
	Lactate	-0.011	-0.029, 0.008	0.264	-0.018	-0.039, 0.003	0.093
	Pyruvate	-0.004	-0.022, 0.015	0.699	-0.017	-0.038, 0.004	0.112
	Citrate	-0.046	-0.064, -0.028	6.26×10^{-7}	0.013	-0.008, 0.034	0.221
Amino acids	**Alanine**	0.027	0.008, 0.045	0.004	-0.019	-0.04, 0.002	0.079
	Glutamine	-0.011	-0.029, 0.007	0.243	0.000	-0.021, 0.02	0.993
	Histidine	0.039	0.02, 0.057	3.59×10^{-5}	-0.008	-0.029, 0.013	0.460
	Isoleucine	0.055	0.036, 0.073	4.16×10^{-9}	-0.044	-0.065, -0.023	4.11×10^{-5}
	Leucine	0.066	0.048, 0.084	9.67×10^{-13}	-0.039	-0.06, -0.018	2.79×10^{-4}
	Valine	0.075	0.057, 0.093	8.23×10^{-16}	-0.050	-0.071, -0.029	2.80×10^{-6}
	Phenylalanine	0.070	0.052, 0.088	8.09×10^{-14}	-0.021	-0.042, 0	0.055
	Tyrosine	0.061	0.042, 0.079	6.92×10^{-11}	-0.028	-0.049, -0.007	0.009
Ketone bodies	**Acetate**	0.029	0.011, 0.048	0.002	-0.032	-0.053, -0.011	0.003
	Acetoacetate	-0.029	-0.047, -0.011	0.001	-0.009	-0.029, 0.012	0.413
	3-hydroxybutyrate	-0.041	-0.06, -0.023	1.12×10^{-5}	-0.008	-0.029, 0.013	0.446
Fluid balance	**Creatinine**	-0.057	-0.074, -0.039	5.57×10^{-10}	0.013	-0.008, 0.033	0.219
	Albumin (signal area)	-0.017	-0.035, 0.001	0.066	0.016	-0.005, 0.036	0.144
Inflammation	**Glycoprotein acetyls**	-0.013	-0.031, 0.005	0.157	0.012	-0.009, 0.032	0.256

Appendix B – First author publications

Richmond RC,* Sharp GC,* Ward ME,* et al. DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework. *Diabetes* 2016; **65**(5): 1231-44.

*joint first authors

Ward ME, McMahon G, St Pourcain B, et al. Genetic variation associated with differential educational attainment in adults has anticipated associations with school performance in children. *PloS one* 2014; **9**(7): e100248.